UNIVERSIDAD CARLOS III DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

INGENIERÍA EN INFORMÁTICA

PROYECTO FIN DE CARRERA

Generación y aplicación de modelos de aprendizaje para la identificación y clasificación de entidades

Autor: José María Rubio Royo-Villanova
Tutora: Mónica Marrero Llinares
Septiembre 2011

Índice

Índice de Tablas

Índice de Ilustraciones

Glosario

TOKEN	Se denomina token a cada una de las palabras o símbolos que aparecen en un fichero.
ATRIBUTO	Característica utilizada para representar algo significativo de un elemento.
POS-TAGGER	Herramienta que se utiliza para obtener las categorías gramaticales de cada una de los tokens de un fichero.
CORPUS DE ENTRENAMIENTO	Fichero o ficheros correctamente etiquetados utilizados para el desarrollo de un modelo de aprendizaje.
CORPUS DE TEST	Fichero o ficheros correctamente etiquetados que se utilizan para realizar las pruebas de los modelos obtenidos a partir de un fichero de entrenamiento.
MVC	Modelo-Vista-Controlador. Patrón de arquitectura de software que separa los datos, la interfaz y la lógica de control de una aplicación.
EXTENSIBLE MARKUP LANGUAGE (XML)	Lenguaje de Marcado Extensible. Subconjunto de SGML que proporciona un método universal para describir e intercambiar información independiente de aplicaciones y proveedores.

1. Introducción

Debido al gran crecimiento de información digital y a su imposibilidad de poder manejarla de forma sencilla, surge la idea de intentar conocer de qué trata un determinado texto. Para intentar conseguir esto, teniendo en cuenta que la investigación en el área de la Extracción y Recuperación de Información está creciendo del mismo modo, nos valemos de una subsección de dicho área para realizar herramientas que lo hagan posible. La subsección a la que nos referimos es la de Reconocimiento y Clasificación de Entidades Etiquetadas (del inglés Named Entity Recognition and Classification, NERC).

Debido a que comprender un texto con estas herramientas no es posible, hablando de una comprensión total, éstas suelen formar parte de herramientas más complejas dedicadas a la Recuperación de Información y a la Búsqueda de Respuestas. De este modo, el manejo de la información se hace más factible para las personas sin necesidad de revisar manualmente ingentes cantidades de información. Lo que se busca con este tipo de herramientas es responder a preguntas del tipo ¿quién?, ¿cuándo?, ¿dónde?

Para ocuparse de este tema, según se comentará más adelante, existen una serie de proyectos y conferencias que se dedican a resolver estos problemas y así intentar llegar a una comprensión lo más amplia posible de un texto. Lo que se quiere conseguir es una herramienta independiente del idioma y del ámbito del texto a analizar (jurídico, deportivo,…). Para ayudar a conseguir esto, se utilizan diversas técnicas de aprendizaje (supervisado, semisupervisado y no supervisado) basadas en listados de entidades, modelos matemáticos y otros recursos que hacen que se obtengan muy buenos resultados.

Aunque por lo general, las herramientas detectan unos pocos tipos de entidades, cada vez más se están desarrollando aplicaciones que tienen en cuenta un número mayor de tipos subdividiendo los tipos principales en subtipos y añadiendo tipos diferentes (como pueden ser numéricos, fechas,…).

Dichas herramientas se basan en atributos implícitos en las palabras de un texto y en otros atributos obtenidos a través de estudios, utilización de herramientas externas o por la aplicación de diferentes modelos de manera secuencial, que dan más información de la palabra que de otra forma no se podrían conocer.

1.1. Objetivos

Este proyecto se ha realizado en base al cumplimiento de diversos objetivos. Además del desarrollo de modelos de identificación y clasificación, gran parte de los objetivos están relacionados con la utilización de diversos recursos externos.

Se podría decir que el objetivo principal de este proyecto, es el desarrollo de una aplicación NER. Puede parecer que no exista ninguna novedad con respecto al resto de aplicaciones orientadas a este tema, pero la aplicación que se desarrollará utilizará un analizador morfo-sintáctico propio para obtener la información gramatical de cada token del fichero a analizar.

Dependiendo del tipo de herramienta a desarrollar en cuanto a los recursos que utiliza, algunas herramientas utilizan listados para ayudar en la identificación y clasificación de las entidades buscadas. Este es el caso de la aplicación a desarrollar, ya que utilizará unos listados obtenidos de la ontología Yago.

Esta herramienta implementará varios modelos de identificación y clasificación, para lo cual utiliza otra herramienta externa (Weka) y para comprobar la calidad de dichos modelos, se apoya en el corpus del idioma inglés utilizado en la conferencia CoNLL ’03. Las entidades que se buscarán serán las de tipo persona, localización, organización y miscelánea.

También se quiere comprobar cuánto influye el entorno de una palabra en que dicha palabra sea una entidad. Para esto, se realizará un estudio contextual con el fichero de entrenamiento del corpus utilizado y se añadirán como atributos al modelo los tokens más indicativos.

Otro aspecto a comprobar, será la implicación de un modelo de identificación concreto, en la correcta clasificación de las entidades. Es decir, se quiere comprobar cuánto influye aplicar un modelo de identificación concreto a la hora de clasificar las entidades. Para ello, el modelo de clasificación se realizará habiendo aplicado dos modelos de identificación diferentes: el peor y el mejor, en cuanto a número de entidades identificadas correctamente.

Al desarrollarse una aplicación que proporciona información acerca de un texto concreto, se quiere que la salida sea utilizable por otras aplicaciones, de modo que se

generará un fichero en formato XML con los porcentajes de acierto de cada modelo aplicado.

1.2. Estructura del proyecto

En este capítulo se ha realizado una breve introducción al proyecto, informando al lector acerca del tema que se va a tratar a lo largo de este documento. En el capítulo dos, "Estado del Arte", se realizará una descripción de la situación actual en el desarrollo de aplicaciones para el reconocimiento de entidades y cómo ha avanzado en los últimos años las capacidades de las mismas. En el capítulo tres, "Herramientas Utilizadas", se describirán todas las herramientas externas (de terceros) que se han utilizado para llevar a cabo el desarrollo del proyecto. El uso de dichas herramientas eran requisitos del sistema. En el capítulo cuatro, "Desarrollo del Proyecto", se mostrará el análisis y diseño de la aplicación así como los problemas encontrados a lo largo de su desarrollo y cómo se resolvieron. También se especificarán las pruebas realizadas para comprobar el funcionamiento de la aplicación. En el capítulo cinco, "Experimentación", se describirán todos los modelos realizados, sus características y los resultados de los mismos a la hora de aplicarlos a los ficheros de entrenamiento y prueba. En el capítulo seis, "Resultados", se analizarán los resultados obtenidos en el capítulo anterior y la relación entre los resultados y los parámetros de cada aproximación. Finaliza el proyecto con un apartado dedicado a las "Conclusiones", donde se explicarán las conclusiones a las que se ha llegado realizando esta aplicación y la experiencia personal que se ha tenido realizando este proyecto; y el capítulo, "Líneas Futuras", que se dedica a mostrar posibles modificaciones para una nueva versión de la aplicación.

2. Estado del Arte

En este apartado, se darán definiciones básicas y útiles a lo largo de este documento y se comentará la evolución del tema tratado.

2.1. Qué es NER

El término "entidad de nombre" (named entity- NE) se estableció en la sexta conferencia MUC (Message Understanding Conference) (MUC-6 Program Committee 1995) (Ralph Grishman & Sundheim 1996). Estas conferencias que comenzaron en 1987, estaban patrocinadas por la agencia norteamericana de defensa DARPA y su fin era la Extracción de Información (IE) de textos no estructurados. En la sexta conferencia MUC, debido a la importancia de identificar información (personas, organismos y localizaciones, en principio) automáticamente, se define NER como el reconocimiento de entidades de nombre y comienza a ser un campo en contínuo crecimiento dentro de la IE.

Aún no se ha llegado a ningún acuerdo para decidir qué abarcan las "entidades de nombre". Por ejemplo, en las conferencias MUC además de los nombres de persona, los organismos y las localizaciones, también incluyen las cantidades y las entidades temporales. Sin embargo, en las conferencias CoNLL-2002/3, se consideran las tres primeras y se añade un nuevo tipo denominado Miscelánea que incluye otros nombres propios de distinto tipo. Por otro lado, las conferencias ACE, añaden más categorías como armas e instalaciones así como consideran la separación de las expresiones temporales como una tarea independiente.

Por lo general, las herramientas NER suelen identificar los tipos de entidades persona, organismo y localización, aunque además pueden identificar otras categorías aunque no suele ser frecuente ya que es necesario adaptar la herramienta con corpus de entrenamiento con esos tipos de entidades anotados.

2.2. Tipos de aprendizaje

Existen diferentes tipos de herramientas NER en cuanto a las técnicas y recursos que utilizan. Existen tres tipos de técnicas: supervisadas, semisupervisadas y no supervisadas. A continuación se comentan los distintos tipos [2]:

Aprendizaje Supervisado

Los sistemas de aprendizaje supervisado, también conocidos habitualmente como clasificadores, se basan en la creación de reglas a partir de ejemplos positivos y negativos anotados en un corpus. La idea general es que un experto anota los fragmentos de texto que servirán como un corpus de documentos de entrenamiento, y el sistema de aprendizaje generalizará a partir de estos ejemplos reglas que puedan ser aplicadas a instancias no conocidas por el sistema (Moens 2006).

Los algoritmos utilizados en este tipo de aprendizaje incluyen Modelos Ocultos de Markov (Hidden Markov Models, HMM), Árboles de Decisión (Decision Trees), Modelos de Máxima Entropía (Maximum Entropy Models, ME), Máquinas de Vectores de Soporte (Support Vector Machines, SVM), y Campos Condicionales Aleatorios (Conditional Random Fields, CRF).

Por lo general, cuanto mayor sea el número de entidades etiquetadas, más efectivo y mayor rendimiento tendrá el aprendizaje supervisado.

El aprendizaje supervisado suele constar de dos pasos:

- Aplicar el algoritmo a un conjunto de datos que ya se encuentran etiquetados (ficheros de entrenamiento y de test).
- Evaluación de los resultados obtenidos. Esta evaluación suele realizarse mediante dos métricas:
    - Precision: número total de entidades encontradas.
    - Recall: de todas las entidades encontradas, cuáles realmente son del tipo en cuestión.
    - F-Measure: la función para medir el rendimiento es:

$$\boldsymbol{F = \frac{2 * precision * recall}{precision + recall}}$$

Aprendizaje No Supervisado

Este tipo de aprendizaje, al contrario que el anterior, no se basa en un conocimiento previo por lo que habitualmente utilizan métodos de agrupamiento (clustering) para realizar la clasificación, o recursos externos (como WorNet), patrones léxicos y estadísticas en base a corpus anotados.

Aprendizaje Semi supervisado

Se trata de un aprendizaje que se encuentra entre el aprendizaje supervisado y el no supervisado. Este tipo de aprendizaje realiza la tarea de clasificación partiendo de un pequeño conjunto de entidades etiquetadas y de un gran corpus sin etiquetar. Una de las técnicas más conocidas en la actualidad es la llamada "bootstrapping" que partiendo de ejemplos anotados identifica patrones sintácticos. Con estos patrones, identifica nuevos ejemplos. Por lo general, este tipo de aprendizaje utiliza patrones léxicos y expresiones regulares para poder clasificar nuevas entidades en relación al corpus etiquetado del que parte.

Un ejemplo de este tipo de aprendizaje, puede ser el de S. Brin (1998) que utiliza expresiones regulares para implementar características léxicas y así generar listas de libros y sus autores. Este algoritmo se basa en la idea de que normalmente, todas las páginas albergadas en un mismo sitio web (website) tienen el mismo formato. Esto hace que con expresiones regulares o restricciones sencillas se puedan detectar nuevos pares libro-autor. Otro ejemplo es la herramienta KnowItAll (Etzioni et al, 2005)

2.3. Principales competiciones en Reconocimiento de Entidades de Nombre

A continuación se van a comentar tres de las más importantes competiciones en el ámbito de NER.

Message Understanding Conferences (MUC)

Mucho ha evolucionado esta conferencia desde que en 1987 comenzara con aspecto exploratorio y todavía sin ningún tipo de evaluación formal de los resultados obtenidos. Según han pasado las ediciones, ha ido aumentando la complejidad de estas conferencias y de las técnicas utilizadas para extraer la información. En las dos primeras ediciones, se daba información acerca de eventos y temas militares a identificar en un texto y se debían rellenar plantillas con la información solicitada. En la segunda edición, también se desarrollaron las medidas de evaluación, recall y precision.

En las ediciones tercera y cuarta, se trata información sobre ataques terroristas y se utilizan plantillas más complejas que en las ediciones anteriores.

MUC-5 representa un gran salto en cuanto a la complejidad de las tareas. Lo más destacable de esta edición, fue el uso de plantillas anidadas en lugar de un único registro en la base de datos con muchos atributos, debido a que en el caso de que en un ataque

terrorista hubiera más de una persona involucrada, sería muy complicado obtener la información de una manera óptima. Al utilizar plantillas anidadas, una plantilla tendría la información del evento y apuntaría a una lista de plantillas, una por cada participante en el evento.

Como en la MUC-5, las tareas habían resultado bastante complejas y habían supuesto un gran esfuerzo por parte del gobierno que tenía que preparar los datos de entrenamiento y los de prueba; y por parte de los asistentes, que trabajaron durante largo tiempo para adaptar sus herramientas a dichas tareas, se decidió concretar una reunión para definir los objetivos iniciales de la MUC-6:

- Tareas a corto plazo para que la información fuera utilizable lo antes posible y demostrar la independencia del dominio. Para conseguir esto, estableció la tarea de entidades de nombre agrupando los nombres de personas, organizaciones y localizaciones geográficas en un texto. Posteriormente, añadieron entidades de tiempo, divisas y expresiones porcetuales.
- Conseguir unos sistemas de IE más portables. Para ello, se decidió que las tareas de IE tendrían plantillas relativamente simples (más como en MUC-2 que como en MUC-5)
- Desarrollar medidas de entendimiento. Desarrollaron tres tareas más para conseguir una medición mejor, debido a que el sistema se basaba principalmente en la comprobación mediante patrones y no era suficiente para conseguir elaborar mecanismos para un conocimiento más profundo.

Decir que una vez definidas las tareas anteriores e implementadas las soluciones, se obtuvieron unas tasas muy elevadas de precision y recall, generalmente superiores al 90%. Concretamente, el mejor sistema obtuvo un 96% de recall y un 97% de precision. Aunque hay que tener en cuenta el número limitado de textos en el conjunto de pruebas (todos extraídos de The Wall Street Journal), los resultados son excelentes.

Conference on Computational Natural Language Learning (CoNLL)

Las conferencias CoNLL se llevan realizando desde el año 1999 gracias a un grupo de aprendizaje del lenguaje natural de la ACL. El tema de la conferencia varía según el año, pero en los año 2002 y 2003, estuvo dedicada al reconocimiento de entidades en diversos lenguajes (español y holandés en 2002, alemán e inglés en 2003) (Sang 2002)

(Sang & Meulder 2003). En estas conferencias, como también se ha comentado anteriormente, se reconocen entidades de tipo persona, organismo, localización y miscelánea. La función de evaluación es F-Measure (β=1). Los corpus de entrenamiento y de test proceden de artículos periodísticos y son de libre acceso (Ilustración 1).

English data	Articles	Sentences	Tokens
Training set	946	14,987	203,621
Development set	216	3,466	51,362
Test set	231	3,684	46,435

English data	LOC	MISC	ORG	PER
Training set	7140	3438	6321	6600
Development set	1837	922	1341	1842
Test set	1668	702	1661	1617

German data	Articles	Sentences	Tokens
Training set	553	12,705	206,931
Development set	201	3,068	51,444
Test set	155	3,160	51,943

German data	LOC	MISC	ORG	PER
Training set	4363	2288	2427	2773
Development set	1181	1010	1241	1401
Test set	1035	670	773	1195

Ilustración 1. Corpus utilizado en CoNLL 2003

En esta edición, participaron 16 sistemas, con medias cercanas al 90% (Ilustración 2) en el mejor de los casos. Todas las herramientas utilizaron técnicas de aprendizaje automático.

English test	Precision	Recall	$F_{\beta=1}$
Florian	88.99%	88.54%	88.76±0.7
Chieu	88.12%	88.51%	88.31±0.7
Klein	85.93%	86.21%	86.07±0.8
Zhang	86.13%	84.88%	85.50±0.9
Carreras (a)	84.05%	85.96%	85.00±0.8
Curran	84.29%	85.50%	84.89±0.9
Mayfield	84.45%	84.90%	84.67±1.0
Carreras (b)	85.81%	82.84%	84.30±0.9
McCallum	84.52%	83.55%	84.04±0.9
Bender	84.68%	83.18%	83.92±1.0
Munro	80.87%	84.21%	82.50±1.0
Wu	82.02%	81.39%	81.70±0.9
Whitelaw	81.60%	78.05%	79.78±1.0
Hendrickx	76.33%	80.17%	78.20±1.0
De Meulder	75.84%	78.13%	76.97±1.2
Hammerton	69.09%	53.26%	60.15±1.3
Baseline	71.91%	50.90%	59.61±1.2

German test	Precision	Recall	$F_{\beta=1}$
Florian	83.87%	63.71%	72.41±1.3
Klein	80.38%	65.04%	71.90±1.2
Zhang	82.00%	63.03%	71.27±1.5
Mayfield	75.97%	64.82%	69.96±1.4
Carreras (a)	75.47%	63.82%	69.15±1.3
Bender	74.82%	63.82%	68.88±1.3
Curran	75.61%	62.46%	68.41±1.4
McCallum	75.97%	61.72%	68.11±1.4
Munro	69.37%	66.21%	67.75±1.4
Carreras (b)	77.83%	58.02%	66.48±1.5
Wu	75.20%	59.35%	66.34±1.3
Chieu	76.83%	57.34%	65.67±1.4
Hendrickx	71.15%	56.55%	63.02±1.4
De Meulder	63.93%	51.86%	57.27±1.6
Whitelaw	71.05%	44.11%	54.43±1.4
Hammerton	63.49%	38.25%	47.74±1.5
Baseline	31.86%	28.89%	30.30±1.3

Ilustración 2. Resultados sobre corpus de test en CoNLL 2003

Automatic Content Extraction (ACE)

El objetivo de este programa, que comenzó en el año 2000 y tuvo su primera competición en 2001, es el desarrollo de tecnología de extracción automática de contenidos que pueda procesar el lenguaje natural en una gran variedad de fuentes, como agentes de noticias, conversaciones y blogs. Se centra en la identificación de entidades, relaciones y eventos. Considera los siguientes tipos de entidades: persona, organización, localización, instalación, entidad geo-política, vehículos y armas. Para cada entidad, se tiene en cuenta el tipo propiamente dicho, el subtipo, la clase y todas las veces que se menciona esa entidad en el texto. Debido a que se deben tener en

cuenta esos parámetros, la forma de evaluar no es tan sencilla como en las otras conferencias. Se encuentra en 4 idiomas: inglés, chino, árabe y español.

En la Tabla 3, se muestran las puntuaciones obtenidas sobre los corpus en inglés en la tarea de reconocimiento de entidades en la competición del año 2008.

Site	Overall	Broadcast Conversations	Broadcast News	Meetings	Newswire	Telephone	Usenet	Weblogs
IBM	50.8%	44.6%	37.7%	-11.9%	58.1%	26.1%	25.5%	51.0%
BBN Technologies	52.6%	42.0%	36.9%	-44.2%	61.3%	22.1%	31.1%	54.8%
Fudan University	-17.6%	-45.1%	-43.3%	-441.2%	9.0%	-197.0%	-48.4%	4.4%
Pontificia Universidade Catolica do Rio de Janeiro, Genesis Institute (Cortex Intelligence)	-46.3%	-54.7%	-21.1%	-57.6%	-64.9%	-18.6%	-48.8%	10.1%
Fondazione Bruno Kessler	-90.0%	-148.1%	-98.8%	-404.6%	-63.8%	-436.2%	-83.0%	-43.5%
AU-KBC Research Center	-269.1%	-340.0%	-279.7%	-911.4%	-188.3%	-999.9*%	-279.6%	-177.4%

Tabla 1. Resultados de ACE'08 para el reconocimiento de entidades en inglés

3. Herramientas y Recursos utilizados

3.1. Yago

Debido a que la aplicación utiliza lenguaje supervisado para identificar y clasificar los tokens de un texto, es habitual el uso de listados de entidades de los tipos que se quieran clasificar. Estos listados se pueden obtener de diferentes maneras. Para el desarrollo de esta aplicación se han utilizado listados procedentes de Yago.

Yago es una base de conocimiento contruída a partir de entidades y relaciones entre las mismas obtenidas de WordNet, Wikipedia y GeoNames. Posee millones de entidades así como decenas de millones de relaciones entre las entidades. Pero Como se puede ver en la siguiente ilustración, yago no es un simple listado de entidades y relaciones entre las mismas, sino que es un complejo sistema que se realimenta continuamente. Se puede dividir en tres partes [1]:

- A.- Esta parte se encarga de obtener artículos de la Wikipedia. Extrae la información de dichos artículos y, debido a que puede haber información inconsistente, utiliza la taxonomía de Wordnet para verificar la consistencia de los datos.
- B.- Las verificaciones que realiza son de tres tipos:
    - De jerarquía: si la información no encaja en el sistema de clases, la desecha y garantiza la consistencia.
    - De tipos: define los tipos de las relaciones y elimina relaciones ambiguas.
    - De restricciones en las relaciones: se realiza inferencia y se eliminan relaciones erróneas.
- C.- Debido a que es necesario mantener la información actualizada, se realimenta obteniendo nueva información de otros recursos web (textos, artículos, noticias, blogs), y revisa la información de Wikipedia y WordNet. También realiza hipótesis y utiliza herramientas de procesamiento del lenguaje natural y valida la información obtenida.

Yago colabora en numerosos proyectos, véase la Ilustración 1. Por ejemplo, contribuye con sus entidades a la ontología UMBEL, se ha fusionado con la

ontología SUMO en el proyecto YAGO-SUMO, se ha exportado su tecnología a Freebase, forma parte de DBpedia y participa en otros muchos proyectos.

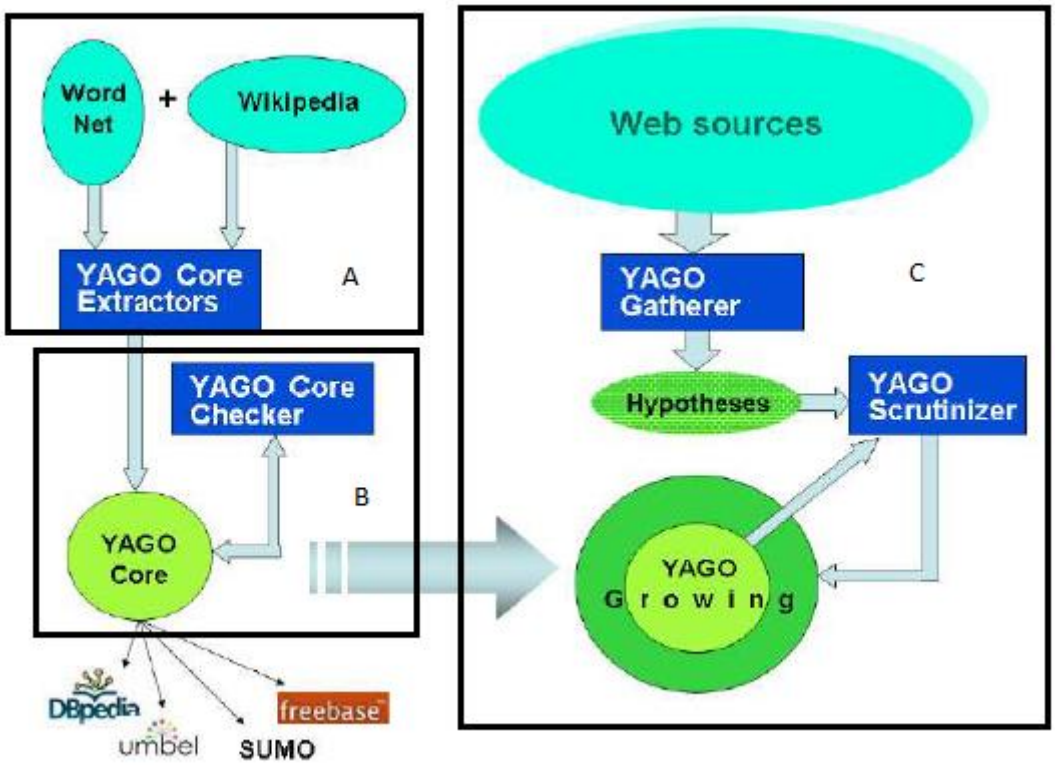

Ilustración 3. Funcionamiento ontología Yago

3.2. Analizador morfo-sintáctico

Se trata de un software desarrollado en el grupo que procesa textos en lenguaje natural, separando los párrafos en palabras o tokens. Además proporciona la categoría gramatical de cada palabra o conjunto de palabras, una normalización de la misma y la posición absoluta y relativa en el texto. También ofrece diversos candidatos para cada palabra (en el caso de que existan varios).

El uso de un analizador de este tipo fue necesario porque en las aplicaciones de aprendizaje automático, es frecuente considerar la categoría gramatical de la palabra como un atributo muy importante.

A continuación se puede ver un ejemplo de lo comentado:

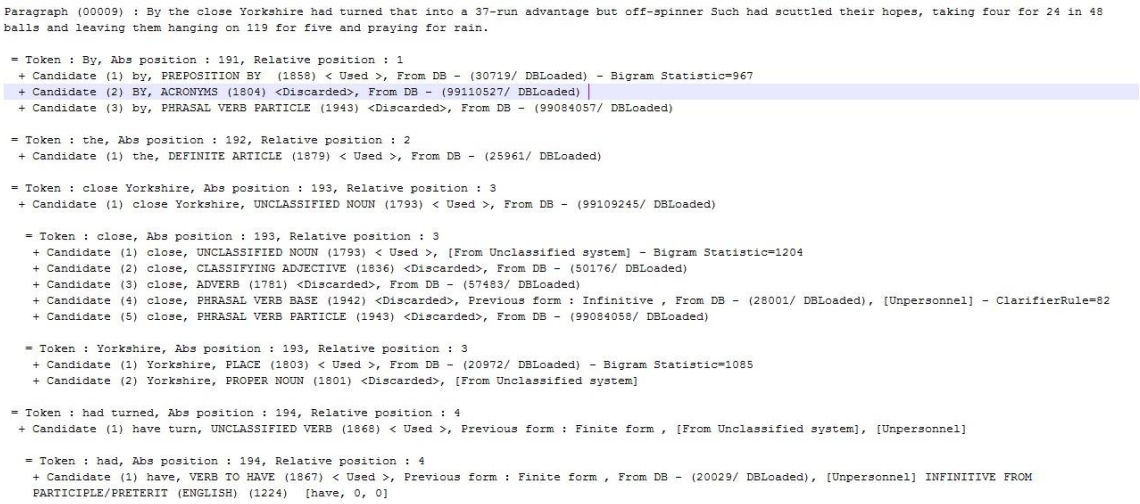

Ilustración 4. Formato LOG Indexer

En la figura anterior, que es un fragmento del fichero de salida del indexer, se puede ver cómo el indexer va descomponiendo la frase palabra por palabra y una vez llegado a lo que para él es un elemento atómico, muestra los diferentes candidatos que ha barajado a la hora de examinar el texto. Obviamente, sólo puede indicar un candidatos de entre los posibles como mejor candidato (Candidate (1)), los demás son descartados.

3.3. Weka

Es una herramienta desarrollada en JAVA y que está orientada al aprendizaje automático y a la minería de datos. Posee gran número de algoritmos.

Este software se ha utilizado para obtener los modelos de identificación y clasificación que se han implementado en la aplicación. Para esto, se han realizado varias pruebas con diversos algoritmos, que se explican a continuación.

3.3.1. J48

Es la implementación de una versión posterior y mejorada del algoritmo C4.5, que es la última versión pública de esta familia de algoritmos (Árboles de Decisión) antes de que apareciera la implementación comercial C5.0. Los árboles de decisión son una forma clásica para representar la información y ofrece una forma rápida y eficaz de expresar las estructuras de datos.

El algoritmo J48 consigue solventar algunos problemas de sus antecesores [4]:

- Evita el sobreajuste (overfitting): no incluye ramas con datos demasiado específicos. Para conseguir esto, implementa nuevas reglas:
    - Pre-poda: una rama para de crecer cuando la información es poco fiable.

    - Post-poda: tomar un árbol de decisión completo y desechar las partes con datos poco fiables.
- Reduce el error de la poda, por lo que se obtiene el subárbol más pequeño con el mismo resultado.
- Permite atributos continuos (con un número infinito de valores): se crea un atributo discreto para comprobar el atributo continuo.
- Permite atributos con muchos valores: modifica la función de medida adecuándola al problema actual.
- Permite atributos con diferentes unidades de medida: reemplaza la función de medida.
- Permite valores de atributos no conocidos: se le asigna el valor más común del atributo según otros ejemplos o se le asigna una probabilidad.
- Añade una regla de post-poda: mediante esta regla convierte el árbol a un conjunto equivalente de reglas. Poda cada regla independientemente y ordena el resultado por uso.

En la imagen siguiente se puede ver una parte de la salida de este algoritmo en Weka:

```
J48 pruned tree
------------------

completoEnListado = si
|   case = upper: entity (4334.0/10.0)
|   case = lower
|   |   numeroPalabrasListado <= 1
|   |   |   categoria_ant = PREPOSITION
|   |   |   |   categoria_sig = PREPOSITION
|   |   |   |   |   en_listado_ant = si: entity (12.0/2.0)
|   |   |   |   |   en_listado_ant = no: noentity (2.0)
|   |   |   |   categoria_sig = ABBREVIATION: entity (0.0)
|   |   |   |   categoria_sig = VERB: entity (12.0/2.0)
|   |   |   |   categoria_sig = ACRONYMS: entity (0.0)
|   |   |   |   categoria_sig = ADJECTIVE: entity (12.0)
|   |   |   |   categoria_sig = ADVERB: entity (1.0)
|   |   |   |   categoria_sig = INVARIANT: noentity (5.0/1.0)
|   |   |   |   categoria_sig = SYMBOL
```

Ilustración 5. Salida algoritmo J48.

3.3.2. PART

Este algoritmo genera una lista de reglas de decisión. Utiliza la misma técnica que el algoritmo anterior, de modo que construye subárboles y posteriormente, convierte el camino desde la raíz hasta la hoja en una regla, como se puede ver en la Ilustración 4.

```
PART decision list
------------------

case = lower AND
de = no AND
bury = no AND
stoke = no AND
completoEnListado = no AND
beat = no AND
categoria = PRONOUN AND
en_listado = si AND
categoria_sig = VERB: noentity (2889.0)

case = lower AND
de = no AND
bury = no AND
stoke = no AND
completoEnListado = no AND
beat = no AND
categoria = DETERMINER AND
cent = no AND
en_listado = si: noentity (14166.0/64.0)

case = lower AND
de = no AND
bury = no AND
stoke = no AND
completoEnListado = no AND
beat = no AND
categoria = PREPOSITION: noentity (15812.0/108.0)
```

Ilustración 6. Salida algoritmo PART

3.4. Corpus

El corpus con el que se ha trabajado es el corpus en inglés utilizado en ConLL 03. Se ha obtenido del corpus de Reuters y está basado en artículos entre agosto de 1996 y agosto de 1997. Dicho corpus consta de un fichero de entrenamiento y dos ficheros de prueba. El fichero de entrenamiento tiene información obtenida de 946 artículos y tiene 203.621 tokens. En esos tokens, existen: 7140 entidades de tipo LOC, 3438 entidades de tipo MISC, 6321 entidades de tipo ORG y 6600 entidades de tipo PER. Los ficheros de prueba contienen información de 231 artículos y existen 46.435 tokens. De esos tokens, 1668 son entidades de tipo LOC, 702 son entidades de tipo MISC, 1661 son entidades de tipo ORG y 1617 son entidades de tipo PER.

Como su propio nombre indica, el fichero de entrenamiento se utiliza para entrenar el modelo y una vez construido se utilizan los ficheros de prueba para comprobar la validez del modelo construido.

Los ficheros de corpus están formateados de tal manera que en cada línea hay una palabra y cada línea consta de 4 columnas (los corpus en alemán constan de 5

columnas). A continuación se puede ver un pequeño fragmento del corpus de entrenamiento [5]:

```
U.N.        NNP  I-NP  I-ORG
official    NN   I-NP  O
Ekeus       NNP  I-NP  I-PER
heads       VBZ  I-VP  O
for         IN   I-PP  O
Baghdad     NNP  I-NP  I-LOC
.           .    O     O
```

Cada columna del ejemplo anterior se describe a continuación:

- la palabra
- la categoría gramatical de la palabra
- etiqueta dada por un divisor de textos (text chunker)
- la entidad a la que corresponde esa palabra

Las palabras etiquetadas con una O no se tienen cuenta para el estudio ya que no forman parte de ninguna entidad. Las etiquetas del tipo B-XXX, se refieren a una palabra que indica el comienzo de una entidad y las etiquetas I-XXX indican que la palabra forma parte de la entidad. Este tipo de notación se denomina BIO.

A continuación se detalla lo que se encuentra englobado dentro de cada categoría:

- Localizaciones:
    - Carreteras: calles y autopistas.
    - **trajectories**
    - Regiones: pueblos, ciudades, provincias, países, continentes, diócesis, parroquias.
    - Estructuras: puentes, puertos, presas.
    - Localizaciones naturales: montañas, cordilleras, bosques, ríos, pozos, campos, valles, jardines, reservas naturales, huertos, playas, parques nacionales.
    - Espacios públicos: plazas, teatros de ópera, museos, escuelas, supermercados, aeropuertos, estaciones, piscinas, hospitales, polideportivos, centros de juventud, parques, ayuntamientos, teatros, cines, galerías, campings, plataformas de lanzamiento de la NASA, universidades, bibliotecas, iglesias, centros médicos, aparcamientos, playgrounds, cementerios.
    - Espacios comerciales: farmacias, bares, restaurantes, almacenes, albergues, hoteles, polígonos industriales, discotecas, auditorios.
    - Edificios de varios: casas, monasterios, orfanatos, fábricas, cuarteles militares, castillos, asilos, torres, residencias, habitaciones, vicarías, courtyards.
    - Sitios abstractos: el mundo libre, por ejemplo.

- Miscelánea:
    - Palabras que forman parte de localizaciones, organizaciones, miscelánea o personas.
    - Adjetivos y otras palabras derivadas de palabras que son localizaciones, organizaciones, miscelánea o persona.
    - Religiones
    - Ideologías políticas
    - Nacionalidades
    - Idiomas
    - Programas
    - Eventos: conferencias, festivales, competiciones deportivas, conferencias, fiestas, conciertos.
    - Guerras
    - Nombres relacionados con el deporte (league tables, ligas, copas)
    - Títulos de libros, canciones, películas, cuentos, discos, programas de televisión, musicales.
    - Eslóganes
    - Periodos de tiempo
    - Tipos de objetos (no marcas): tipos de coche, aviones, motocicletas.
- Organizaciones:
    - Compañías: agencias de noticias, estudios, bancos, cooperativas, fabricantes, mercados de valores.
    - Subdivisiones de compañías (newsrooms)
    - Marcas
    - Movimientos políticos: partidos políticos, organizaciones terroristas, organismos gubernamentales (ministros, alcaldes, tribunales, uniones políticas de países (Naciones Unidas).
    - Publicaciones: revistas, periódicos, diarios.
    - Compañías musicales: grupos, compañías de ópera, orquestas, coros.
    - Organizaciones públicas: colegios, universidades, organizaciones benéficas.
    - Otros grupos de personas: clubes deportivos, equipos, asociaciones, companies teatrales, órdenes religiosas, organizaciones juveniles.
- Personas
    - Nombre y apellidos de personas
    - Animales y personajes ficticios
    - Alias

4. Desarrollo del proyecto

En este capítulo, se mostrarán todas las fases del proyecto que se han seguido para el desarrollo de la aplicación.

4.1. Análisis

En esta fase se definirán los requisitos solicitados y posteriormente, una vez se haya implementado la aplicación y se realicen las pruebas, se comprobará su cumplimiento.

4.1.1. Especificación de Requisitos

La información que se detallará para especificar los requisitos es la siguiente:

- **Identificador**: conjunto de caracteres que identifican unívocamente a cada requisito. Se divide en dos partes unidas por un guión. La primera indica el tipo de requisito, mientras que la segunda parte es el número que identifica el requisito.
- **Nombre**: nombre significativo del requisito.
- **Descripción**: explicación detallada del requisito.
- **Necesidad:** nivel de necesidad dentro del sistema. Puede tomar siguientes los valores: Esencial, Conveniente u Opcional.
- **Estabilidad**: posibilidad de que el requisito esté sujeto a cambios a lo largo del desarrollo de la aplicación. Puede tomar los valores: Estable o No estable.
- **Prioridad**: nivel de prioridad del requisito. Puede tomar siguientes los valores: Alta, Media o Baja.
- **Fuente**: Procedencia del requisito.

A continuación se especifican los requisitos:

IDENTIFICADOR	R-001
NOMBRE	Importar fichero
DESCRIPCIÓN	La aplicación debe ser capaz de importar un fichero con un determinado formato. Este formato es el formato de salida del Indexer. A partir de ahora, se denominará **fichero de entrada**.
NECESIDAD	☑ Esencial ☐ Conveniente ☐ Opcional
ESTABILIDAD	☑ Estable ☐ No estable
PRIORIDAD	☑ Alta ☐ Media ☐ Baja

FUENTE	Cliente.

Tabla 2. R-001 Importar fichero

IDENTIFICADOR	R-002
NOMBRE	Listados de Yago.
DESCRIPCIÓN	La aplicación debe ser capaz de usar los listados de Yago para comprobar si los tokens del fichero de entrada pertenecen al listado de alguna categoría concreta.
NECESIDAD	☑ Esencial ☐ Conveniente ☐ Opcional
ESTABILIDAD	☑ Estable ☐ No estable
PRIORIDAD	☑ Alta ☐ Media ☐ Baja
FUENTE	Cliente

Tabla 3. R-002 Listados

IDENTIFICADOR	R-003
NOMBRE	Modelo de identificación
DESCRIPCIÓN	La aplicación debe implementar como mínimo un modelo de identificación para poder realizar la identificación.
NECESIDAD	☑ Esencial ☐ Conveniente ☐ Opcional
ESTABILIDAD	☐ Estable ☑ No estable
PRIORIDAD	☑ Alta ☐ Media ☐ Baja
FUENTE	Cliente

Tabla 4. R-003 Modelo de identificación

IDENTIFICADOR	R-004
NOMBRE	Modelo de clasificación
DESCRIPCIÓN	La aplicación debe implementar como mínimo un modelo de clasificación para poder realizar la clasificación.
NECESIDAD	☑ Esencial ☐ Conveniente ☐ Opcional
ESTABILIDAD	☐ Estable ☑ No estable
PRIORIDAD	☑ Alta ☐ Media ☐ Baja

FUENTE	Cliente

Tabla 5. R-004 Modelo de clasificación

IDENTIFICADOR	R-005
NOMBRE	Preprocesado
DESCRIPCIÓN	Se realizará un preprocesado de los datos de entrada para obtener información necesaria para la ejecución de los modelos. Esta información es: pertenencia a los ficheros de listados, completitud en los mismos y número de palabras que forman parte de la entidad encontrada.
NECESIDAD	☐ Esencial ☑ Conveniente ☐ Opcional
ESTABILIDAD	☐ Estable ☑ No estable
PRIORIDAD	☑ Alta ☐ Media ☐ Baja
FUENTE	Cliente

Tabla 6. R-005 Preprocesado

IDENTIFICADOR	R-006
NOMBRE	Nombres y verbos frecuentes
DESCRIPCIÓN	Será necesario realizar un estudio para ver cuáles son los nombres y verbos que más se repiten y comprobar, a la hora de hacer los modelos, si influyen dichas palabras en mejorar el reconocimiento.
NECESIDAD	☐ Esencial ☐ Conveniente ☑ Opcional
ESTABILIDAD	☑ Estable ☐ No estable
PRIORIDAD	☐ Alta ☑ Media ☐ Baja
FUENTE	Cliente

Tabla 7. R-006 Estudio

IDENTIFICADOR	R-007
NOMBRE	Identificación
DESCRIPCIÓN	La aplicación debe ser capaz de realizar la identificación de las palabras indicando si una palabra es o no entidad así como la probabilidad de serlo.
NECESIDAD	☑ Esencial ☐ Conveniente ☐ Opcional

ESTABILIDAD	☑ Estable ☐ No estable
PRIORIDAD	☑ Alta ☐ Media ☐ Baja
FUENTE	Cliente

Tabla 8. R-007 Identificación

IDENTIFICADOR	R-008
NOMBRE	Clasificación
DESCRIPCIÓN	La aplicación debe ser capaz de realizar la clasificación de las palabras indicando el tipo de entidad que es (PER, LOG, ORG, MISC, NA) así como la probabilidad de serlo.
NECESIDAD	☑ Esencial ☐ Conveniente ☐ Opcional
ESTABILIDAD	☑ Estable ☐ No estable
PRIORIDAD	☑ Alta ☐ Media ☐ Baja
FUENTE	Cliente

Tabla 9. R-008 Clasificación

IDENTIFICADOR	R-009
NOMBRE	Ejecución automática
DESCRIPCIÓN	Todo el proceso debe poderse ejecutar de manera automática. Únicamente el usuario deberá pulsar el botón de ejecución tras seleccionar un fichero de texto como entrada.
NECESIDAD	☐ Esencial ☑ Conveniente ☐ Opcional
ESTABILIDAD	☑ Estable ☐ No estable
PRIORIDAD	☑ Alta ☐ Media ☐ Baja
FUENTE	Cliente

Tabla 10. R-009 Ejecución automática

IDENTIFICADOR	R-010
NOMBRE	Fichero de salida

DESCRIPCIÓN	Será necesario que la aplicación genere un fichero en el que se verá la probabilidad de cada token para el modelo de identificación y para el modelo de clasificación.
NECESIDAD	☑ Esencial ☐ Conveniente ☐ Opcional
ESTABILIDAD	☑ Estable ☐ No estable
PRIORIDAD	☑ Alta ☐ Media ☐ Baja
FUENTE	Cliente

Tabla 11. R-010 Fichero de salida

IDENTIFICADOR	R-011
NOMBRE	Directorios de listados de Yago.
DESCRIPCIÓN	La aplicación dará la opción de seleccionar los directorios en los que se encuentran los listado de yago.
NECESIDAD	☐ Esencial ☑ Conveniente ☐ Opcional
ESTABILIDAD	☑ Estable ☐ No estable
PRIORIDAD	☐ Alta ☑ Media ☐ Baja
FUENTE	Cliente

Tabla 12. R-011 Directorios de listados

IDENTIFICADOR	R-012
NOMBRE	Atributos básicos para la clasificación.
DESCRIPCIÓN	Los atributos básicos que se utilizarán para realizar la clasificación de las entidades serán: case {upper,lower,title,mixed} contieneDigitos {si,no} contieneSimbolos {si,no} categoria categoria_sig categoria_ant en_listado_per {si,no} en_listado_org {si,no} en_listado_loc {si,no} en_listado_misc {si,no} en_listado_per_ant {si,no} en_listado_org_ant {si,no} en_listado_loc_ant {si,no} en_listado_misc_ant {si,no} en_listado_per_sig {si,no}

	en_listado_org_sig {si,no} en_listado_loc_sig {si,no} en_listado_misc_sig {si,no} completoEnListado {si,no}
NECESIDAD	☑ Esencial ☐ Conveniente ☐ Opcional
ESTABILIDAD	☐ Estable ☑ No estable
PRIORIDAD	☑ Alta ☐ Media ☐ Baja
FUENTE	Cliente

Tabla 13. R-012 Atributos básicos para la clasificación

IDENTIFICADOR	R-013
NOMBRE	Atributos básicos para la identificación.
DESCRIPCIÓN	Los atributos básicos que se utilizarán para realizar la identificación de las entidades serán: case {upper,lower,title,mixed} contieneDigitos {si,no} contieneSimbolos {si,no} categoria categoria_sig categoria_ant en_listado {si,no} en_listado_ant {si,no} en_listado_sig {si,no} completoEnListado {si,no} numeroPalabrasListado numeric
NECESIDAD	☑ Esencial ☐ Conveniente ☐ Opcional
ESTABILIDAD	☐ Estable ☑ No estable
PRIORIDAD	☑ Alta ☐ Media ☐ Baja
FUENTE	Cliente

Tabla 14. R-013 Atributos básicos para la identificación

4.1.2. Diagrama de casos de uso

Entendiendo Caso de Uso como una funcionalidad completa de la herramienta, a continuación se muestra el diagrama de casos de uso de la aplicación (Ilustración 7):

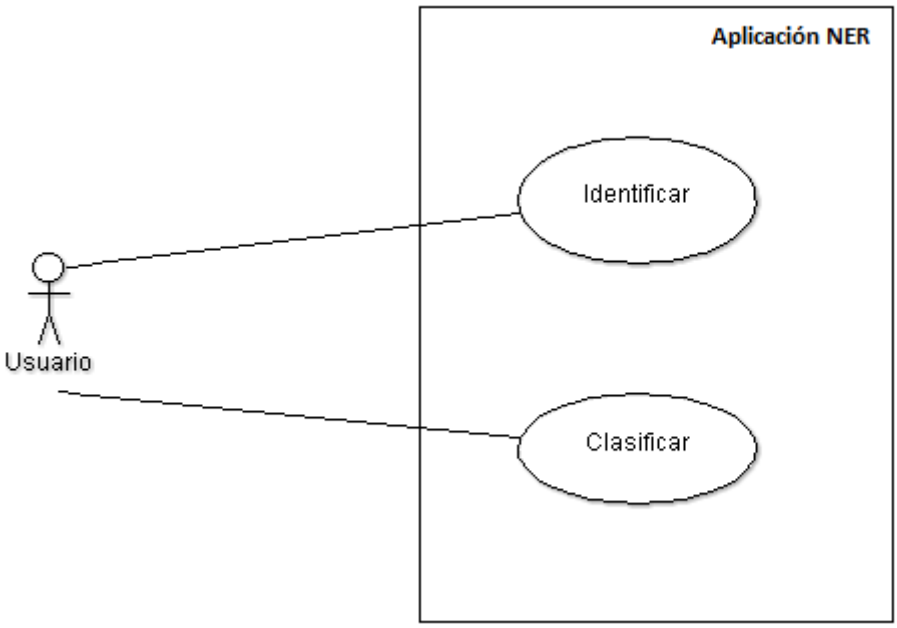

Ilustración 7. Diagrama de Casos de Uso

La información que se detallará para especificar los casos de uso es la siguiente:

- **Identificador:** conjunto de caracteres que identifican unívocamente a cada caso de uso.
- **Nombre:** nombre significativo del caso de uso.
- **Descripción:** explicación detallada del caso de uso.
- **Actores:** personas que intervienen en la ejecución del caso de uso.
- **Precondiciones:** estado del sistema que se tiene que cumplir para que el caso de uso se pueda ejecutar.
- **Postcondiciones:** estado del sistema una vez ejecutado el caso de uso.
- **Escenario:** pasos que se deben seguir para la correcta ejecución del caso de uso.

A continuación se detallan los casos de uso del diagrama anterior:

IDENTIFICADOR	CU-01
NOMBRE	Identificar.
DESCRIPCIÓN	La aplicación obtendrá la información de aplicar el modelo de identificación.
ACTORES	Usuario
PRECONDICIONES	Preprocesado realizado.

POSTCONDICIONES	Identificación realizada.
ESCENARIO	Iniciar aplicación. Cargar listados. Cargar fichero de entrada. Parsear fichero entrada. Preprocesado de identificación. Acceder al menú "Modelo Identificación" y pulsar la opción "Identificar".

Tabla 15. CU-05 Identificar

IDENTIFICADOR	CU-02
NOMBRE	Clasificar.
DESCRIPCIÓN	La aplicación obtendrá la información de aplicar el modelo de clasificación.
ACTORES	Usuario
PRECONDICIONES	Identificación realizada.
POSTCONDICIONES	Clasificación realizada. Obtención del fichero de salida
ESCENARIO	Iniciar aplicación. Cargar listados. Cargar fichero entrada. Parsear fichero entrada. Preprocesado de identificación. Identificar. Acceder al menú "Modelo Clasificación" y pulsar la opción "Clasificar".

Tabla 16. CU-06 Clasificar

Al tratarse de casos de uso que representan una funcionalidad completa, resultan bastante complejos aunque ambos casos de uso tiene el mismo orden de ejecución. Se va a realizar un diagrama de actividad del caso de uso C01-Identificar para aclarar los pasos a seguir para completar su ejecución:

Ilustración 8. Diagrama de Actividad de CU-01 Identificar

Como se puede ver en la Ilustración 8, hay varias actividades que conforman el caso de uso por lo que se van a comentar:

- **Iniciar aplicación**

  Obviamente, la primera acción para llevar a cabo la identificación, será iniciar la aplicación.

- **Cargar listados**

  La siguiente tarea a realizar es la carga de los ficheros de listados en memoria para su posterior utilización. Además, aparte de añadir las entidades tal cual se encuentran en los ficheros de listados y, para intentar abarcar más posibles opciones, se añaden las palabras individuales que forman dicha entidad así como grupos de palabras incrementales desde el inicio hasta el fin de la entidad. Esto sólo para las entidades de Persona y Localización.

- **Cargar fichero de entrada**

Es necesario tener un fichero de entrada para identificar sus entidades por lo que se deberá cargar en la aplicación.

- **Parsear fichero de entrada**

  Se extrae la información útil del fichero de entrada cargado y se almacena en memoria en la estructura con la que se trabajará a lo largo de la aplicación.

- **Preprocesado**

  Básicamente se comprueba si los tokens del fichero de entrada se encuentran en los ficheros de listados, tanto individualmente como por grupos de palabras adyacentes en el texto. Como consecuencia de este preprocesado, se consigue saber en qué listados se encuentra una palabra, si está completa en los mismos o si varios tokens consecutivos del fichero de entrada se encuentran en alguno de los listados, formando así una única entidad.

- **Identificar**

  Para realizar una de las funciones principales, se seleccionará el Modelo de Identificación a ejecutar y se realizará la identificación.

- **Mostrar información**

  Una vez ejecutada la identificación de las entidades, se mostrará información de los resultados. Debido a que si el fichero es demasiado grande, no sería práctico mostrarlo todo en pantalla, únicamente se muestran unos tokens para comprobar el resultado de la identificación.

4.1.3. Atributos Utilizados

En este apartado se van a explicar los atributos utilizados a la hora de obtener los modelos, identificación y clasificación.

- **Atributos presentes en ambos modelos:**
  - Case: los valores posibles para este atributo son: upper (la palabra está en mayúsculas), lower (la palabra está en minúsculas), title (la primera letra de la palabra es la única mayúscula), mixed (palabra formada por letras mayúsculas y minúsculas).
  - ContieneDigitos: atributo de tipo lógico. Indica que la palabra actual está formada por algún dígito.
  - ContieneSimbolos: atributo lógico que indica si la palabra posee símbolos distintos a letras y números.

- Categoría: este atributo puede tomar uno de los siguientes valores: PREPOSITION,ABBREVIATION,VERB,ACRONYMS,ADJECTIVE, ADVERB,INVARIANT,SYMBOL,DETERMINER,NOUN,PRONOUN ,'LLCHART VARIABLE',NONE,'PHRASAL VERB PARTICLE',SOFTWARE. Indica la categoría gramatical de la palabra.
- Categoria_sig: los valores posibles son los mismos que los del atributo anterior. Este atributo indica la categoría gramatical de la palabra siguiente a la que se está analizando.
- Categoria_ant: puede tener un valor de entre los indicados en el atributo 'categoría'. Indica la categoría gramatical de la palabra anterior a la que se está analizando.
- CompletoEnListado: atributo de tipo lógico (si, no) que indica si la palabra o grupo de palabras se encuentran textualmente en los listados.
- NumeroPalabrasListado: atributo de tipo numérico que indica el número de palabras del fichero de entrada que forman parte de la entidad de los listados.
- Atributos del análisis: Véase el apartado 4.3.2.

- **Atributos del modelo de Identificación:**
  - En_listado: atributo de tipo lógico que indica si la palabra actual se encuentra en los listados.
  - En_listado_ant: atributo de tipo lógico que indica si la palabra anterior a la actual se encuentra en los listados.
  - En_listado_sig: atributo de tipo lógico que indica si la palabra siguiente a la actual se encuentra en los listados.
- **Atributos del modelo de Clasificación:**
  - En_listado_per: atributo de tipo lógico que indica que la palabra actual se encuentra en el/los listados de Personas.
  - En_listado_org: atributo de tipo lógico que indica que la palabra actual se encuentra en el/los listados de Organizaciones.
  - En_listado_loc: atributo de tipo lógico que indica que la palabra actual se encuentra en el/los listados de Localizaciones.
  - En_listado_misc: atributo de tipo lógico que indica que la palabra actual se encuentra en el/los listados de Miscelánea.

- En_listado_per_ant: atributo de tipo lógico que indica que la palabra anterior a la actual se encuentra en el/los listados de Personas.
- En_listado_org_ant: atributo de tipo lógico que indica que la palabra anterior a la actual se encuentra en el/los listados de Organizaciones.
- En_listado_loc_ant: atributo de tipo lógico que indica que la palabra anterior a la actual se encuentra en el/los listados de Localizaciones.
- En_listado_misc_ant: atributo de tipo lógico que indica que la palabra anterior a la actual se encuentra en el/los listados de Miscelánea.
- En_listado_per_sig: atributo de tipo lógico que indica que la palabra siguiente a la actual se encuentra en el/los listados de Personas.
- En_listado_org_sig: atributo de tipo lógico que indica que la palabra siguiente a la actual se encuentra en el/los listados de Organizaciones.
- En_listado_loc_sig: atributo de tipo lógico que indica que la palabra siguiente a la actual se encuentra en el/los listados de Localizaciones.
- En_listado_misc_sig: atributo de tipo lógico que indica que la palabra siguiente a la actual se encuentra en el/los listados de Miscelánea.
- AciertosIdentificacion: atributo de tipo numérico que indica el número de aciertos obtenidos para esa palabra al aplicar el modelo de Identificación.
- FallosIdentificacion: atributo de tipo numérico que indica el número de fallos obtenidos para esa palabra al aplicar el modelo de Identificación.
- Es_entidad_clasificacion: atributo de tipo lógico que indica si, según el modelo de identificación, esa palabra es entidad o no.
- Es_entidad_corpus: atributo de tipo lógico que indica si, según la información del corpus original, esa palabra es entidad o no.

Dependiendo del modelo se utilizarán unos u otros atributos y dependiendo de los atributos utilizados, los resultados serán diferentes. A la hora de describir cada modelo se indicarán los atributos que se utilizaron para su obtención.

4.2. Diseño

A la hora de realizar el diseño de la aplicación, se quería que cada parte de la misma fuera lo más independiente posible para así favorecer su modularidad. De esta forma, la aplicación se puede dividir en tres módulos:

- Un módulo se encarga de la interfaz con la que interactúa el usuario.
- Otro módulo se encarga de la gestión de los datos.
- El último módulo se encarga de la interacción entre los dos módulos anteriores.

Este diseño se asemeja al típico patrón Modelo-Vista-Controlador (MVC). A continuación se muestra una imagen de este patrón tan conocido:

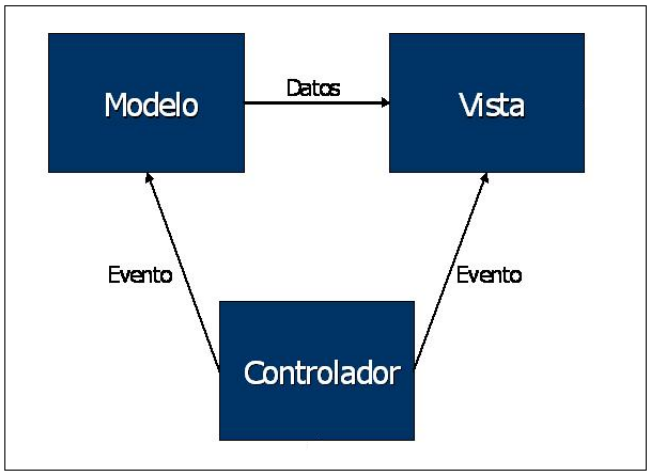

Ilustración 9. Diagrama MVC

Con este tipo de diseño, se logra la independencia de la implementación de cada elemento, de modo que cada módulo no sabe cómo está implementado el resto pero sabe la información que necesita enviar para conseguir la información requerida.

En la siguiente figura se puede ver la organización lógica de la aplicación:

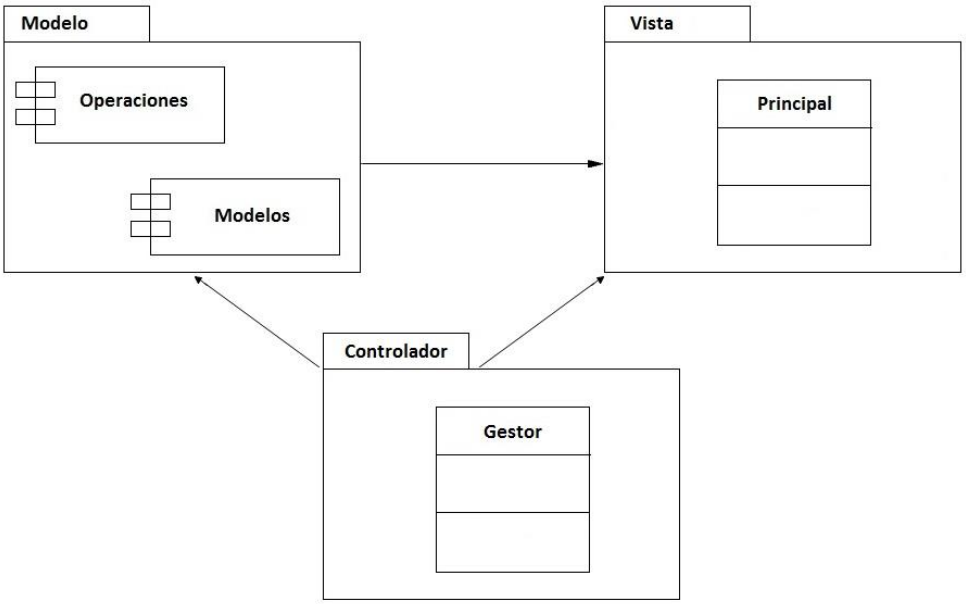

Ilustración 10. Diagrama de componentes del sistema

Según se ha visto en la figura anterior, el sistema se encuentra dividido en distintos módulos según la funcionalidad de cada uno.

- **Vista:** únicamente lo forma la clase Principal, correspondiente al formulario que se muestra al iniciar la aplicación. Gestiona la parte gráfica de la aplicación y la interacción con el usuario mediante la impresión de información acerca del proceso de ejecución.
- **Controlador:** formado por la clase Gestor, que contiene la funcionalidad necesaria para relacionar la Vista con el Modelo. Toda la información pasa a través de él.
- **Modelo:** aquí es donde se encuentra la mayoría de la funcionalidad de la aplicación. Como se puede ver en la figura consta de dos componentes lógicos:
    - **Operaciones:** se ocupa de las operaciones previas a la ejecución de los modelos sobre el fichero de entrada.
    - **Modelos:** agrupa las implementaciones de los modelos de identificación y clasificación que forman parte de la aplicación.

En la ilustración que se muestra a continuación se puede ver el diagrama de clases del sistema:

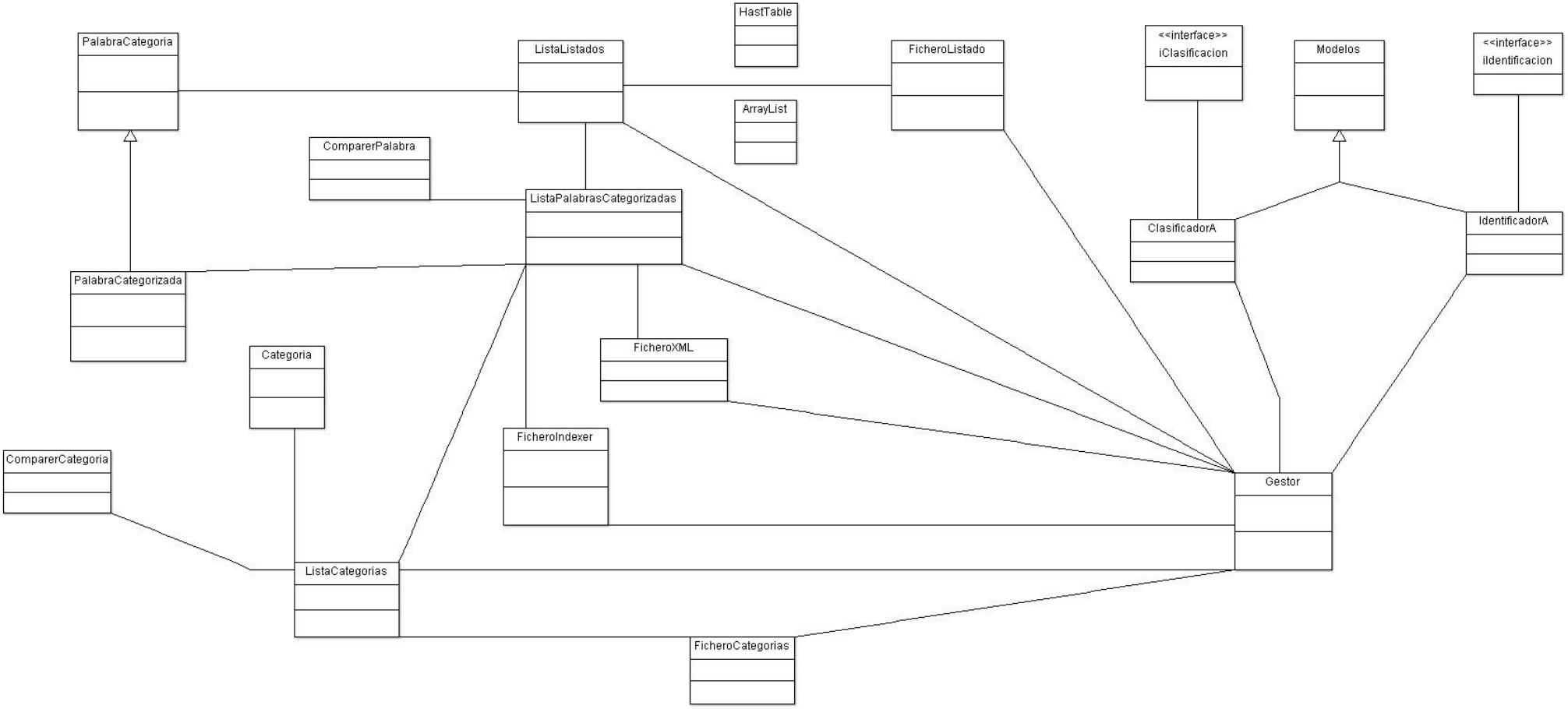

Ilustración 11. Diagrama de Clases

4.2.1. Vista

La capa Vista la forma el formulario principal de la aplicación. En el caso de ser una aplicación más compleja, la formarían todos los formularios existentes. Mediante el formulario principal, el usuario de la aplicación realiza todas las operaciones definidas.

Según el patrón seguido, la capa Vista se relacionará con la capa Controlador mediante el envío de información y eventos.

4.2.2. Controlador

Encargado del correcto funcionamiento de la aplicación, debido a que todos los flujos de comunicación de la aplicación pasan a través de él y de este modo gestiona la comunicación entre la Vista y el Modelo.

Como se ha dicho antes, la única clase que forma el Controlador es la clase Gestor. A continuación, se puede ver su especificación y su descripción:

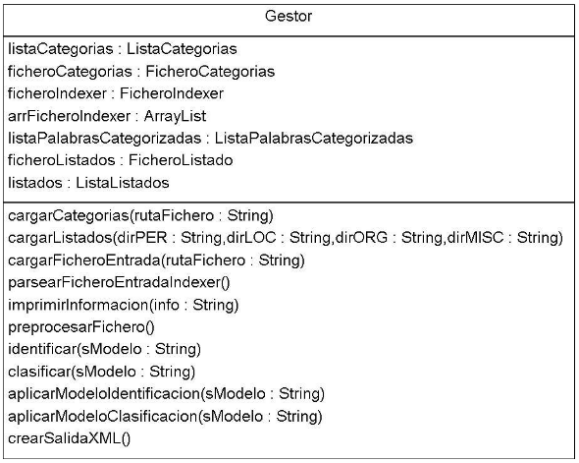

Ilustración 12. Clase Gestor

NOMBRE	Gestor.
DESCRIPCIÓN	Clase que gestionará el tráfico de información entre todos los elementos de la aplicación así como, el orden de ejecución de cada paso de la aplicación.

Tabla 17. Descripción de la clase Gestor

4.2.3. Modelo

El Modelo se encuentra formado por todos los elementos relacionados con el modelo de negocio de la aplicación. Como se dijo antes, el Modelo se encuentra dividido lógicamente en dos partes: Operaciones y Modelos. A continuación se muestra la información de cada uno de ellos:

- Operaciones: formado por todas las clases necesarias para tratar el fichero de entrada y llegar a una estructura con la información necesaria para aplicar el Modelo de Identificación. Las clases que lo forman son:

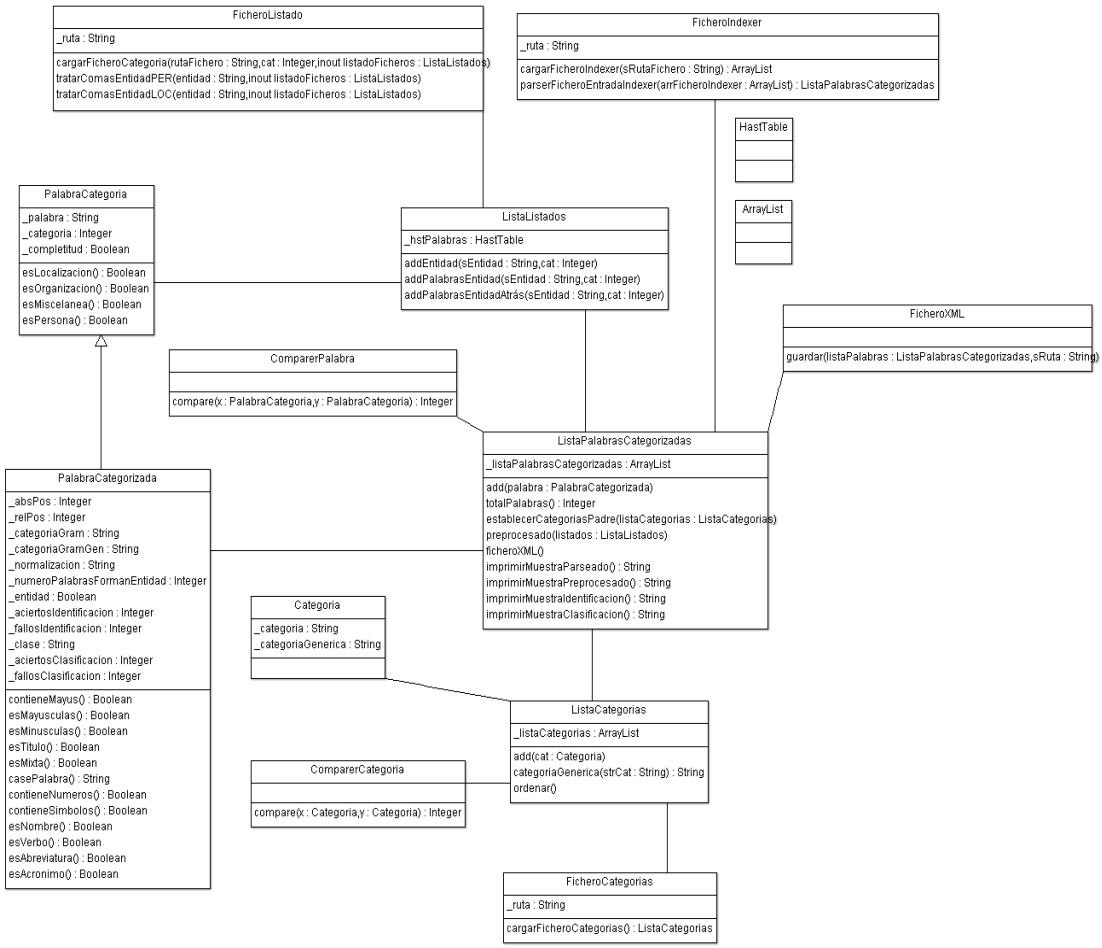

Ilustración 13. Modelo de clases relacionado con las operaciones

A continuación se describen brevemente las clases mostradas en el diagrama anterior:

NOMBRE	PalabraCategoría

DESCRIPCIÓN	Clase que almacenará una entidad de los ficheros de listados, tanto la palabra en sí como el/los listado/s en los que está presente.

Tabla 18. Descripción de la clase PalabraCategoria

NOMBRE	PalabraCategorizada
DESCRIPCIÓN	Clase en la que se almacenará un token del fichero de entrada y al que se irá añadiendo información a medida que avance la ejecución de la aplicación.

Tabla 19. Descripción de la clase PalabraCategorizada

NOMBRE	ListaListados
DESCRIPCIÓN	Clase que almacenará todas las entidades contenidas en los ficheros de listados.

Tabla 20. Descripción de la clase ListaListados

NOMBRE	Categoria
DESCRIPCIÓN	Clase que almacenará una categoría gramtical y su respectiva categoría general.

Tabla 21. Descripción de la clase Categoria

NOMBRE	ListaPalabrasCategorizadas
DESCRIPCIÓN	Clase que contendrá todos los tokens del fichero de entrada así como su información (normalización, posición relativa,…) y con la que se trabajará durante la ejecución de la aplicación.

Tabla 22. Clase ListaPalabrasCategorizadas

NOMBRE	ListaCategorias
DESCRIPCIÓN	Clase en la que se almacenarán todas las categorías gramaticales con las que se va a trabajar así como, las categorías generales de cada una de ellas.

Tabla 23. Descripción de la clase ListaCategorias

NOMBRE	FicheroIndexer
DESCRIPCIÓN	Clase encargada de leer el fichero de entrada a la aplicación y de almacenar la información en la estructura correcta.

Tabla 24. Descripción de la clase FicheroIndexer

NOMBRE	FicheroListados
DESCRIPCIÓN	Clase encargada de leer y cargar en memoria todos los listados con los que se va a trabajar.

Tabla 25. Descripción de la clase FicheroListados

NOMBRE	FicheroCategorias
DESCRIPCIÓN	Clase encargada de leer y cargar en memoria todas las categorías gramaticales con las que se va a trabajar. En el mismo fichero se encuentran las categorías gramaticales genéricas a las que pertenece cada categoría gramatical.

Tabla 26. Descripción clase FicheroCategorias

NOMBRE	FicheroXML
DESCRIPCIÓN	Clase que se encarga de crear el fichero de salida de la aplicación en formato XML y de toda la información que se incluye en el mismo

Tabla 27. Descripción de la clase FicheroXML

NOMBRE	ComparerPalabra
DESCRIPCIÓN	Clase encargada de comparar dos palabras a la hora de hacer búsquedas binarias en la estructura de ListaPalabrasCategorizadas (que contiene los tokens del fichero de entrada).

Tabla 28. Descripción de la clase ComparerPalabra

NOMBRE	ComparerCategoria
DESCRIPCIÓN	Clase encargada de comparar dos categorías gramaticales cuando se realiza la ordenación de la estructura en la que éstas se almacenan (ListaCategorias)

Tabla 29. Descripción de la clase ComparerCategoria

- Modelos: en esta parte, se agrupan las clases relacionadas con la implementación de los modelos así como de la funcionalidad necesaria durante la Identificación y la Clasificación. En la siguiente imagen, se pueden ver las clases que lo forman:

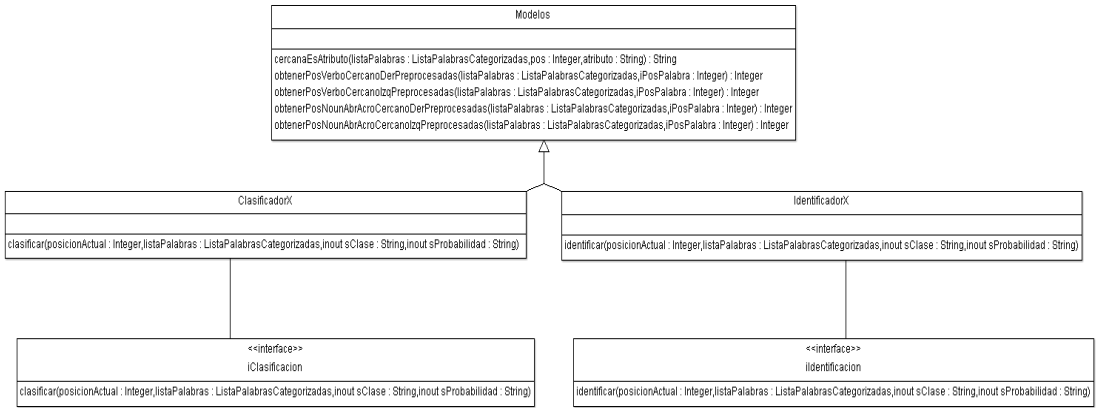

Ilustración 14. Modelo de clases relacionado con los Modelos

A continuación, se realiza una descripción textual de cada una de ellas. Únicamente hay que remarcar que las clase IdentificadorX y ClasificadorX engloban a todas las implementaciones de los modelos de identificación y clasificación en la aplicación, que no se han incluido para dejar una figura más clara, aunque se describan.

NOMBRE	Modelos
DESCRIPCIÓN	Clase que contiene la información común a los diferentes modelos que se han implementado.

Tabla 30. Descripción de la clase Modelos

NOMBRE	IdentificadorA
DESCRIPCIÓN	Clase que contiene la implementación del modelo 7 de identificación y que se encargará de procesar un token del fichero de entrada y proporcionar los datos necesarios.

Tabla 31. Descripción de la clase IdentificadorA

NOMBRE	IdentificadorB
DESCRIPCIÓN	Clase que contiene la implementación del modelo 3 de identificación y que se encargará de procesar un token del fichero de entrada y proporcionar los datos necesarios.

Tabla 32. Descripción de la clase IdentificadorB

NOMBRE	ClasificadorA
DESCRIPCIÓN	Clase que contiene la implementación del modelo 10 de clasificación y que se encargará de procesar un token del fichero de entrada y proporcionar los datos necesarios.

Tabla 33. Descripción de la clase ClasificadorA

NOMBRE	ClasificadorB
DESCRIPCIÓN	Clase que contiene la implementación del modelo 12 de clasificación y que se encargará de procesar un token del fichero de entrada y proporcionar los datos necesarios.

Tabla 34. Descripción de la clase ClasificadorB

4.3. Implementación

En este apartado se va a proceder a explicar todos los pasos seguidos hasta la obtención de la aplicación final. También se comentarán los diferentes modelos obtenidos en el proceso.

4.3.1. Acciones resueltas

Estas acciones se refieren a las llevadas a cabo hasta obtener la aplicación final. Debido a que se debían utilizar herramientas externas, se tuvieron que realizar distintos pasos tanto para ejecutar dichas aplicaciones (formatear la información al formato de entrada de la aplicación) como para comprender su salida (obtener la información útil de la salida de la aplicación). Dichas acciones fueron:

- **Pasar ficheros a texto**

  En el primer paso, debido a que el formato de los ficheros de entrenamiento y de test no era el correcto para poder utilizarlo de entrada a la aplicación externa que se utiliza como analizador gramatical (Indexer), fue necesario convertir a texto plano los ficheros de entrenamiento y test. Para realizar esta operación se probaron dos posibilidades:

  - Como en los ficheros iniciales constan de una palabra por línea, se creó un fichero con cada palabra (o símbolo) separada, es decir, delante y detrás de cada palabra había un espacio (salvo de las de inicio y fin de línea).
  - Para crear un fichero en texto plano correcto, se desarrolló un procedimiento que controlara la posición de los símbolos para ubicarlos correctamente en el texto.

- **Indexer**

  Una vez obtenido el fichero en texto plano, había que ejecutar el Indexer y comprobar con qué configuración funcionaba mejor, ya que dependiendo de la configuración utilizada en el Indexer, la salida sería diferente e influiría en el

resto del proceso. Según las pruebas que se realizaron, el Indexer calificaba de forma distinta a la misma palabra con distintas configuraciones.

- **Parsear log Indexer**

  Como se ha comentado anteriormente, el fichero de salida del Indexer tiene un formato con el que no se podía trabajar por lo que fue necesario desarrollar un procedimiento para parsearlo y así almacenar la información útil para nosotros en la estructura elegida. Esta acción se corresponde con la actividad Parseado del caso de uso CU-01 Identificar.

- **Unir información**

  Debido a que se tenía información útil en dos ficheros distintos, por un lado el fichero del corpus (con la palabra y el tipo de entidad) y por otro el fichero de salida del Indexer (palabra, normalización, categoría gramatical,…), fue necesario realizar una unión de dicha información que no fue tan sencilla como se creía. En el apartado 4.3.3 se pueden ver los problemas encontrados al realizar esta tarea.

- **Ficheros de listados**

  La siguiente tarea a realizar es la carga de los ficheros de listados en memoria para su posterior utilización. Para almacenar esta información se utiliza una tabla hash de modo que la comprobación de si un elemento existe o no es directa. Además, aparte de añadir las entidades tal cual se encuentran en los ficheros de listados y, para intentar abarcar más posibles opciones, se añaden las palabras individuales que forman dicha entidad así como grupos de palabras incrementales desde el inicio hasta el fin de la entidad. Esto sólo para las entidades de Persona y Localización.

- **Preprocesado**

  Esto consiste en comprobar si las palabras que forman el fichero de entrada se encuentran en los ficheros de listados, tanto individualmente como por grupos de palabras adyacentes en el texto. Como consecuencia de este preprocesado, se consigue saber en qué listados se encuentra una palabra, si está completa en los mismos o si varias palabras consecutivas del fichero de entrada se encuentran en alguno de los listados, formando así varias palabras una única entidad.

- **Análisis contextual**

  Véase el apartado 4.3.2.

- **Obtener fichero arff Identificación**

Fue necesario crear los ficheros que utiliza Weka para desarrollar el modelo (entrenamiento y test) a partir de la información que se tenía de cada token.

- **Parsear Modelo Identificación**

  Al igual que la entrada a Weka, el formato de la salida tampoco era utilizable directamente por la aplicación, para lo que se desarrolló un método para parsear el modelo obtenido de Weka y convertirlo a un formato de implementación válido.

- **Obtener fichero arff Clasificación**

  Después de aplicar el modelo de identificación se debe hacer lo mismo para obtener el modelo de clasificación, por lo que se crean nuevos ficheros de entrada a Weka. Los ficheros difieren tanto en los atributos utilizados, en la información que contiene como en el modelo de identificación utilizado.

- **Parsear Modelo Clasificación**

  De la misma forma que se hizo cuando se obtuvo el modelo de identificación, se realiza la conversión de la salida de Weka a una implementación válida.

4.3.2. Análisis contextual

Para obtener un modelo más completo, se pensó realizar un análisis sobre el entorno de las palabras del fichero de entrenamiento. Para ello, se creó un fichero en el que para cada palabra se incluía la siguiente información:

- Palabra: la palabra actual del fichero de entrenamiento.
- Categoría: categoría gramatical de la palabra
- Es_entidad: valor lógico que indica si la palabra era o no entidad.
- Naa_cercano_izq: palabra correspondiente al nombre, abreviatura o acrónimo más cercano a la palabra por la izquierda en la misma frase.
- Clase_naa_izq: tipo de entidad (LOC, PER, ORG, MISC, NA) del nombre, abreviatura o acrónimo más cercano por la izquierda en la misma frase.
- Categoria _naa_izq: categoría gramatical del nombre, abreviatura o acrónimo más cercano a la izquierda de la palabra en la misma frase.
- Naa_cercano_der: palabra correspondiente al nombre, abreviatura o acrónimo más cercano a la palabra por la derecha en la misma frase.
- Clase_naa_der: tipo de entidad (LOC, PER, ORG, MISC, NA) del nombre, abreviatura o acrónimo más cercano por la derecha en la misma frase.

- Categoria _naa_der: categoría gramatical del nombre, abreviatura o acrónimo más cercano a la derecha de la palabra en la misma frase.
- V_cercano_izq: palabra correspondiente al verbo más cercano a la palabra por la izquierda en la misma frase.
- Clase_v_izq: tipo de entidad (LOC, PER, ORG, MISC, NA) del verbo más cercano por la izquierda en la misma frase.
- Categoria _v_izq: categoría gramatical verbo más cercano a la izquierda de la palabra en la misma frase. Este atributo siempre tendrá el valor 'VERB', pero se mantiene para tener la misma estructura que con los nombres, abreviaturas y acrónimos.
- V_cercano_der: palabra correspondiente al verbo más cercano a la palabra por la derecha en la misma frase.
- Clase_v_der: tipo de entidad (LOC, PER, ORG, MISC, NA) del verbo más cercano por la derecha en la misma frase.
- Categoria_v_der: categoría gramatical verbo más cercano a la derecha de la palabra en la misma frase. Este atributo siempre tendrá el valor 'VERB', pero se mantiene para tener la misma estructura que con los nombres, abreviaturas y acrónimos.

Toda esta información se almacenó en una base de datos para facilitar la obtención de la información que se quería obtener a partir de ésta. Para esto se ejecutaron una serie de consultas con las que se obtuvo para cada palabra cuya categoría gramatical genérica fuera 'NOUN' o 'VERB', los siguientes datos:

- Número de veces que aparece la palabra a la izquierda de cualquier entidad.
- Número de veces que aparece la palabra a la derecha de cualquiera entidad.
- Número de veces que aparece la palabra a la izquierda de cualquiera palabra que no sea entidad.
- Número de veces que aparece la palabra a la derecha de cualquiera palabra que no sea entidad.

A partir de estos datos, para cada palabra se realizó el sumatorio de veces que aparecía más cerca por la izquierda o por la derecha de una 'entidad' (E), y el sumatorio de veces que aparecía más cerca por la izquierda o por la derecha de una 'no entidad' (N). Una vez obtenidos los sumatorios, se obtuvo:

- T = E + F

- P = ABS(E - F) / T

Con ello y dependiendo de los valores P:

- Si P > 0,5 entonces V = P * log10(T)
- Si no V = 0

y

- VP = V / MAX(V)

Debido a que se tenían demasiadas palabras para poder utilizar, se decidió utilizar únicamente las que tuvieran un VP > 0,6. A continuación se puede ver una tabla con los atributos que se utilizan en los modelos, denominados a partir de ahora Atributos del Análisis. Estos atributos son de tipo lógico e indican son el nombre, abreviatura, acrónimo o verbo más cercano a la palabra del fichero de entrada que se está tratando.

year	people	months	weeks	one	eight
percent	hour	billion	time	six	beat
percentage	cent	halftime	lower	growth	men
day	million	per	passenger	market	days
rate	total	goal	wicket	price	end
tonnes	vs	rupees	km	de	guilder
metres	demand	two	bln	profit	minute
out	period	sector	pct	three	five
figure	up	gross	order	bushel	accept
programme	bury	think	allocate	exercise	weight
import	specify	sell	contribute	provide	lose
compare	fall	show	cooper	bond	close
stoke	nickel	poise	steady	rise	ensure
change	expect	reserve	create	yen	lot
down	share				

Tabla 35. Atributos del Análisis

4.3.3. Problemas encontrados

A lo largo del desarrollo del proyecto, se han ido encontrando problemas que se han ido resolviendo para poder continuar. A continuación se describen los problemas encontrados:

Problemas encontrados al utilizar Yago

Al tratarse de una base tan grande, se dispone de unos ficheros de entidades que ocupan bastante espacio en disco. Debido a esto, la carga de los listados en memoria lleva algo de tiempo con la consiguiente ocupación de memoria, aunque en la actualidad no sea tan crítica la ocupación de memoria.

Debido a que no se ha utilizado Yago de la manera para la que se ha construido, a la hora de revisar los listados, nos dimos cuenta de que no toda la información de cada listado era válida para el entorno de la aplicación, por lo que fue necesario trabajo manual para intentar conseguir unos listados más correctos.

Problemas encontrados al utilizar el analizador morfo-sintáctico

Normalmente cuando se utiliza una herramienta externa que no fue creada para integrarse con la aplicación que está desarrollando una persona, surgen diversas situaciones que hay que solventar si se quiere continuar con su uso. Dichas situaciones se comentan a continuación:

- No siendo un problema sino una fase que hay que superar cuando se utiliza una herramienta nueva, es intentar conocer su funcionamiento. Aunque únicamente se utiliza una parte de la herramienta, se tuvo que probar diferentes configuraciones y comparar los resultados para decidir con qué configuración se obtenía el mejor resultado.
- Debido a que el formato de salida de la aplicación no era entendible por la aplicación en desarrollo, se tuvo que realizar un parseador para obtener un fichero entendible por la aplicación con toda la información del indexador.
- El resultado del proceso de tokenización de esta herramienta y de la aplicación para el etiquetado de los textos utilizados era diferente. El analizador utilizado:
    - Puede considerar como un token más de una palabra si es considerada palabra compuesta (en los ficheros con los que se prueba, cada token es una única palabra).

- Elimina algunos puntos de final de línea (que son considerados como tokens en los ficheros etiquetados).
- Cambia de sitio tokens.
- Elimina las contracciones de las formas verbales inglesas.

Por todo esto, se tuvo que hacer un fichero para indicar que los tokens de ambos ficheros no coincidían y hacer los cambios a mano en el parseado de entrada, añadiendo los tokens eliminados y poniendo en el sitio correcto los tokens movidos. Para esto, fue necesario imprimir la información a fichero y luego hacer un procedimiento para la carga de un fichero con el formato que se le dio (este fichero ya estaba parseado). Para resolver el problema de las contracciones se desarrolló un procedimiento que tenía en cuenta que el analizador podía haber deshecho la contracción.

4.3.4. Principales algoritmos

Para que se pueda entender un poco mejor el funcionamiento del sistema se va a proceder a comentar los procedimientos más importantes:

- Preprocesado: este procedimiento se encarga de añadir la información relacionada con los listados. Para ello se recorre la estructura en la que se encuentra almacenado el fichero de entrada una vez parseado y para cada elemento realiza una búsqueda en los listados que posee. En el caso de que encuentre alguna coincidencia en los listados, irá comprobando la existencia en los listados de la palabra junto con las palabras que le siguen una a una hasta que hubiera un grupo de palabras que no encontrara en los listados. En ese momento, establecería los valores de los atributos para las palabras que formaban la última entidad que se encontró en los listados. En el caso de no encontrarla en los listados, continuaría con las siguientes palabras. Por ejemplo, si tuviéramos la siguiente frase: Pedro Martín roba pan, se buscaría “Pedro” en los listados. Si lo encontrara, buscaría “Pedro Martín”. Si encontrara también esa entidad, buscaría “Pedro Martín roba” y como seguramente no lo encontraría, establecería que la palabra “Pedro” y la palabra “Martín” se encuentran en el listado formando una entidad de 2 palabras y que se ha encontrado en los listados de personas. A continuación se incluye el pseudocódigo de este procedimiento:

```
Mientras no se termine el fichero
    Obtener elemento actual
    Obtener elemento siguiente
    Obtener coincidencias listado
    Si hubo alguna coincidencia
        Completa en listado = true
        Mientras no se termine el fichero y haya coincidencias
            Obtener palabra siguiente (contador + número de palabras que forman la entidad anterior)
            Concatenar la palabra siguiente a la entidad
            Obtener coincidencias listado de la entidad
            Si hubo alguna coincidencia
                Guardar entidad
                Completa en listado = true
                Aumentar el número de palabras que forman la entidad
            Fin si
         Fin mientras
    Fin si

    Si se encontró alguna entidad
        Para cada palabra que forma parte de la entidad hacer
            Obtener la palabra
            Establecer los ficheros de listados en los que se ha encontrado la palabra
            Palabra.Completa en listado = true
            Establecer el número de palabras que forman parte de la entidad
        Fin para
        Posicionar el cursor en la posición adecuada del array
    Fin si
    Inicializar variables
Fin mientras
```

Ilustración 15. Pseudocódigo preprocesado

- AplicarModeloIdentificación: este procedimiento recorre toda la estructura en la que está almacenada la entrada (modificada por el parseado y por el preprocesado) y por cada token de la estructura, ejecuta el Modelo de Identificación (previamente seleccionado). Una vez realizada la identificación del token, se establecen los valores en los atributos del token. En el caso de que no haya fallos en la identificación (si al generar el modelo de identificación, ninguna prueba resultó fallida), únicamente se devuelve el número de aciertos. En caso contrario, se devuelven ambos con el formato "acierto/fallos". A continuación se puede ver el pseudocódigo:

```
Establecer el modelo a ejecutar
Para cada palabra del fichero de entrada
    Ejecutar el modelo de identificación para la palabra actual
    Establecer el atributo 'entidad' (boolean)
    Si hay aciertos y fallos
        Establecer aciertos
        Establecer fallos
    Si no (solo hay aciertos)
        Establecer los aciertos de la palabra
    Fin si
    Inicializar variables
Fin para
```

Ilustración 16. Pseudocódigo AplicarModeloIdentificación

- Modelos: los procedimientos que realizan las operaciones más importantes de la aplicación son los modelos de identificación y clasificación implementados. Dichos procedimientos constan de miles de líneas de código por lo que sería inviable si inclusión en este documento. Por este motivo y para hacerse una idea de cómo son, a continuación se incluirá un pequeño fragmento de pseudocódigo de uno de los modelos implementados:

```
Si la palabra es título (únicamente con la primera letra en mayúsculas)
    Si está completa en los listados
        entity (12368.0/567.0)
    Si no está compelta en los listados
        Si la categoría de la palabra anterior es PREPOSITION
            Si la categoría de la palabra siguiente PREPOSITION
                Si se encuentra en los listados
                    noentity (262.0/76.0)
                Si no
                    entity (10.0)
                Fin si
                Si la categoría de la palabra siguiente es ABBREVIATION
                    noentity (0.0)
                Si la categoría de la palabra siguiente es VERB
                    Si la palabra siguiente se encuentra en los listados
                        Si el número de palabras que forman la entidad es 1
                            entity (176.0/58.0)
                        Si el número de palabras que forman la entidad es mayor que 1
                            Si el número de palabras que forman la entidad es 2
                                noentity (18.0/6.0)
                            Si el número de palabras que forman la entidad es mayor que 2
                                entity (6.0)
                            Fin si
                        Fin si
                    Si la palabra siguiente no se encuentra en los listados
                        noentity (25.0/5.0)
                    Fin si
                Fin si
                Si la categoría de la palabra siguiente es ACRONYMS
                    Si contiene simbolos
                        noentity (4.0/1.0)
                    Si no contiene simbolos
                        entity (8.0)
                    Fin si
                Fin si
                Si la categoría de la palabra siguiente es ADJECTIVE
                    entity (75.0/19.0)
                Si la categoría de la palabra siguiente es ADVERB
                    noentity (33.0/14.0)
                Si la categoría de la palabra siguiente es INVARIANT
                    Si se encuentra en los listados
                        Si contiene símbolos
                            entity (5.0)
                        Si no contiene símbolos
                            noentity (177.0/79.0)
                        Fin si
                ...
```

Ilustración 17. Pseudocódigo parcial Modelo Identificación

4.4. Pruebas

En este apartado se va a proceder a especificar las pruebas que se han realizado para comprobar el correcto funcionamiento de la aplicación. Todas las pruebas que se realizaron se comprobaron de forma manual y fueron superadas con éxito.

NOMBRE	Carga de los listados
DESCRIPCIÓN	En esta prueba se comprobó que se cargaba correctamente la información de los listados en memoria.
RESULTADO ESPERADO	La estructura en la que se almacena la información de los listados (HashTable) contiene toda la información de los ficheros de listado.

Tabla 36. Prueba Carga de los listados

NOMBRE	Carga del fichero de entrada
DESCRIPCIÓN	Prueba en la que se comprobó que se cargaba en memoria el contenido del fichero de entrada (el fichero de entrada es el LOG de salida del Indexer).
RESULTADO ESPERADO	Toda la información del fichero de entrada se almacenó correctamente en memoria para su posterior uso.

Tabla 37. Prueba Carga del fichero de entrada

NOMBRE	Parseo del fichero de entrada
DESCRIPCIÓN	Prueba encargada de mostrar que el parseo del fichero de entrada se realizaba correctamente y la información útil se genera de forma adecuada.
RESULTADO ESPERADO	La información necesaria para trabajar se queda almacenada y organizada correctamente.

Tabla 38. Prueba Parseo del fichero de entrada

NOMBRE	Preprocesado
DESCRIPCIÓN	Prueba encargada de comprobar que el preprocesado de la información de entrada se ejecutaba correctamente.
RESULTADO ESPERADO	La información relativa a la aparición o no de las palabras en los ficheros de listados queda almacenada junto a la demás información del token.

Tabla 39. Prueba Preprocesado

NOMBRE	Ejecución del Modelo de Identificación correcto.
DESCRIPCIÓN	Se realizó una prueba para comprobar que el Modelo de Identificación que se selecciona en el formulario principal es el que se ejecuta a la hora de ir realizar la identificación.
RESULTADO ESPERADO	Se ejecuta el modelo previamente seleccionado.

Tabla 40. Prueba Ejecución del Modelo de Identificación correcto

NOMBRE	Identificación
DESCRIPCIÓN	Comprobar que la información relativa proporcionada al ejecutar el Modelo de Identificación se genera adecuadamente.
RESULTADO ESPERADO	Los atributos relativos a la identificación almacenan la información importante aportada por el Modelo de Identificación.

Tabla 41. Prueba Identificación

NOMBRE	Clasificación
DESCRIPCIÓN	Comprobar que la información relativa proporcionada al ejecutar el Modelo de Clasificación se genera adecuadamente.
RESULTADO ESPERADO	Los atributos relativos a la clasificación almacenan la información importante aportada por el Modelo de Clasificación.

Tabla 42. Prueba Clasificación

NOMBRE	Ejecución contínua.
DESCRIPCIÓN	Prueba realizada para comprobar que funciona la opción de ejecutar la aplicación con un solo click de ratón.
RESULTADO ESPERADO	Se ejecutan todos los pasos y se obtiene el fichero de salida en formato xml con toda la información importante.

Tabla 43. Prueba Ejecución contínua

NOMBRE	Fichero de salida
DESCRIPCIÓN	Prueba en la que se debía comprobar la correcta generación de un fichero en formato xml con la información obtenida durante el proceso.
RESULTADO ESPERADO	Se obtiene un fichero de salida con toda la información.

Tabla 44. Prueba Fichero de salida

NOMBRE	Información de la ejecución
DESCRIPCIÓN	Comprobar que se muestra información relativa a la ejecución de la aplicación
RESULTADO ESPERADO	La información sobre el progreso de la aplicación se muestra en el formulario principal.

Tabla 45. Prueba Información de la ejecución

5. Experimentación

En este apartado se van a comentar los diferentes modelos que se realizaron, tanto de identificación como de clasificación, los atributos y la información de los mismos. También se muestran la parte de los resultados más importante para ayudar a seleccionar posteriormente los más indicativos. Previamente, se describirán los dos métodos de evaluación con los que se ha experimentado para obtener los modelos.

5.1. Métodos de evaluación

A la hora de trabajar con Weka para tratar de obtener un modelo, existen varias opciones para realizar las pruebas. Para la obtención de los modelos se ha probado con dos de ellas. Esas opciones son:

Validación cruzada (Cross-validation)

La validación cruzada es un método estadístico para la evaluación y comparación de algoritmos de aprendizaje mediante la división del conjunto de los datos en dos subconjuntos: uno se utiliza para entrenar el modelo y el otro se utiliza como datos de prueba para validar el modelo. En la validación cruzada típica, los subconjuntos de entrenamiento y validación se deben cruzar en rondas sucesivas de tal manera que cada dato pueda ser validado. La validación cruzada básica es la que tiene k conjuntos de muestra (k-fold cross-validation). Existen otras formas pero o son casos especiales de la anterior o se refieren a ciclos repetidos de la anterior [3].

En la forma básica del algoritmo, los datos se dividen en k conjuntos del mismo tamaño (o semejante). Posteriormente, se llevan a cabo k iteraciones de entrenamiento y validación, teniendo en cuenta que en cada iteración se reserva un conjunto de muestra distinto para la validación mientras que los k-1 conjuntos restantes se utilizan para el aprendizaje.

Este algoritmo, se utiliza para evaluar o comparar algoritmos de aprendizaje de la siguiente manera: en cada iteración, uno o más algoritmos utilizan k-1 subconjuntos de datos para que aprendan uno o más modelos y posteriormente, se comprueban dichos modelos con el subconjunto de datos reservado anteriormente para la validación. El rendimiento de cada algoritmo con cada conjunto de muestra se puede seguir mediante el uso de alguna métrica del rendimiento como puede ser la precisión.

Al finalizar, tendremos k mediciones por cada algoritmo. Para obtener una medida de dichas mediciones se pueden utilizar diferentes metodologías como el promedio. También se podrían utilizar dichas mediciones para realizar hipótesis estadísticas y mostrar que un algoritmo es mejor que otro.

Conjunto de prueba

En el caso de que contemos con un fichero de prueba válido, es decir con los tokens etiquetados de la misma manera que el fichero de entrenamiento, podemos utilizar esta opción para probar el modelo realizado con ese fichero y así comprobar sus tasas de acierto. Este método da mejores resultados que el anterior debido a que son más fiables porque ambos ficheros son diferentes.

A la hora de implementar los algoritmos, se han utilizados los obtenidos utilizando un conjunto de prueba, ya que en el corpus de la conferencia CoNLL ’03 que se utiliza, hay tanto fichero de entrenamiento como ficheros de pruebas. Los resultados entre unos modelos y otros, no eran lo suficientemente significativos como para implementar los que utilizaban cross-validation, ya que, como se ha dicho, los modelos obtenidos con un conjunto de prueba tienen mayor fiabilidad.

5.2. Modelos de Identificación

Antes de implementar un modelo de identificación concreto se obtuvieron varios para comprobar el porcentaje de acierto de cada uno dependiendo de la información y de los atributos utilizados. A la hora de obtener los modelos, según la forma en la que se ha trabajado se necesitan crear 2 ficheros weka (ficheros con extensión arff y según la gramática de weka): el de entrenamiento y el de prueba. Como fichero de entrenamiento se partió del fichero eng_train.list y como fichero de prueba engA.list, ambos del corpus provisto por el cliente.

El primer modelo se hizo para comprobar el correcto funcionamiento y ver la influencia del uso de listados, ya que como se puede ver es el que mayor porcentaje de entidades identificadas correctamente obtiene y está cercano al 100%. Resultado esperado debido a que utiliza sus propios listados.

A continuación se detallan los 9 modelos de identificación diferentes, así como los resultados obtenidos:

- **Modelo 1:** para obtener el fichero de entrenamiento se utilizaron listados obtenidos de sí mismo, al igual que para la obtención del fichero de prueba. Los atributos utilizados para la obtención de este modelo fueron: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado, En_listado_ant, En_listado_sig, CompletoEnListado y NumeroPalabrasListado. Los resultados proporcionados por Weka al generar este modelo son los siguientes:

Correctly Classified Instances	50394	99.1403 %
Incorrectly Classified Instances	437	0.8597 %

Class	Precision	Recall	F-measure
Entity	0,975	0,975	0,975
Noentity	0,995	0,995	0,995

Tabla 46. Resultados Modelo 1 de Identificación

- **Modelo 2:** en este modelo únicamente se utilizaron los listados obtenidos del fichero de entrenamiento, tanto para sí mismo como para el fichero de prueba. Se utilizaron los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado, En_listado_ant, En_listado_sig, CompletoEnListado y NumeroPalabrasListado. Con estas características, los resultados fueron:

Correctly Classified Instances	48249	94.9204 %
Incorrectly Classified Instances	2582	5.0796 %

Class	Precision	Recall	F-measure
Entity	0,973	0,73	0,827
Noentity	0,948	0,996	0,97

Tabla 47. Resultados Modelo 2 de Identificación

- **Modelo 3:** los listados utilizados en este modelo fueron los obtenidos del fichero de entrenamiento. Los atributos que utiliza este modelo son: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado, En_listado_ant, En_listado_sig, CompletoEnListado,

NumeroPalabrasListado y Atributos del modelo. Los resultados obtenidos fueron:

Correctly Classified Instances	48341	95.1014 %
Incorrectly Classified Instances	2490	4.8986 %

Class	Precision	Recall	F-measure
Entity	0,976	0,717	0,834
Noentity	0,945	0,996	0,971

Tabla 48. Resultados Modelo 3 de Identificación

- **Modelo 4:** para comprobar el grado de alcance de la utilización de ficheros a la hora de ir a crear el modelo, en este caso no se utilizó ningún tipo de listado. Los atributos utilizados en este modelo fueron: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, CompletoEnListado y NumeroPalabrasListado. Se obtuvieron los siguientes resultados:

Correctly Classified Instances	48619	95.6483 %
Incorrectly Classified Instances	2212	4.3517 %

Class	Precision	Recall	F-measure
Entity	0,893	0,843	0,868
Noentity	0,969	0,979	0,974

Tabla 49. Resultados Modelo 4 de Identificación

- **Modelo 5:** al igual que en el modelo anterior, para obtener este modelo tampoco se utilizaron listados. Los atributos utilizados en este caso fueron: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, CompletoEnListado, NumeroPalabrasListado y Atributos del Análisis. Los resultados que se obtuvieron fueron los siguientes:

Correctly Classified Instances	48648	95.7054 %
Incorrectly Classified Instances	2183	4.2946 %.

Class	Precision	Recall	F-measure
Entity	0.893	0,847	0,869
Noentity	0,969	0,979	0,974

Tabla 50. Resultados Modelo 5 de Identificación

- **Modelo 6:** para obtener este modelo se utilizaron listados procedentes de la ontología Yago. Los atributos que se utilizaron fueron: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado, En_listado_ant, En_listado_sig, CompletoEnListado y NumeroPalabrasListado. Se obtuvieron los siguientes resultados:

Correctly Classified Instances	48914	96.2287 %
Incorrectly Classified Instances	1917	3.7713 %

Class	Precision	Recall	F-measure
Entity	0,913	0,859	0,885
Noentity	0,972	0,983	0,997

Tabla 51. Resultados Modelo 6 de Identificación

- **Modelo 7:** este modelo utiliza los listados obtenidos de la ontología Yago y se basa en los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado, En_listado_ant, En_listado_sig, CompletoEnListado, NumeroPalabrasListado y Atributos del Análisis. Los resultados de este modelo fueron:

Correctly Classified Instances	48942	96.2838 %
Incorrectly Classified Instances	1889	3.7162 %

Class	Precision	Recall	F-measure
Entity	0,908	0,868	0,888
Noentity	0,973	0,982	0,978

Tabla 52. Resultados Modelo 7 de Identificación

- **Modelo 8:** para este modelo se realizaron algunos cambios en los listados de yago ya que se vio que había entidades que no pertenecían a la entidad del fichero (error cometido al crear los ficheros de listados) y únicamente incluyendo las palabras individuales que forman las entidades y los grupos incrementales para las entidades de tipo PER. Los atributos utilizados fueron: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado, En_listado_ant, En_listado_sig, CompletoEnListado y NumeroPalabrasListado. A continuación se muestran los datos del modelo:

Correctly Classified Instances	48908	96.2169 %

Incorrectly Classified Instances 1923 3.7831 %

Class	Precision	Recall	F-measure
Entity	0,906	0,866	0,886
Noentity	0,973	0,982	0,977

Tabla 53. Resultados Modelo 8 de Identificación

- **Modelo 9:** para este modelo también se realizaron cambios en los ficheros de Yago (los mismos que en el caso anterior) y sólo se incluyeron las palabras individuales que forman las entidades y los grupos incrementales para las entidades de tipo PER. Los atributos utilizados fueron: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado, En_listado_ant, En_listado_sig, CompletoEnListado, NumeroPalabrasListado y Atributos del Análisis. Los resultados obtenidos fueron:

Correctly Classified Instances 48919 96.2385 %
Incorrectly Classified Instances 1912 3.7615 %

Class	Precision	Recall	F-measure
Entity	0,906	0,867	0,886
Noentity	0,973	0,982	0,977

Tabla 54. Resultados Modelo 9 de Identificación

Debido a que tanto el Modelo 1 como el Modelo 2 utilizan listados obtenidos de los ficheros de entrenamiento y prueba, no se tendrán en cuenta a la hora de hablar del modelo que mejor o peor porcentaje obtiene.

5.3. Modelos de Clasificación

Una vez implementados los modelos de identificación en la aplicación, se hizo la misma operación para obtener los modelos de clasificación. Como con el modelo anterior, se hicieron varias pruebas para comprobar la importancia de los atributos utilizados y de la información contenida.

Los 8 primeros modelos se obtuvieron utilizando el modelo de identificación con peor porcentaje en la obtención de entidades identificadas (Modelo 3, ya que como hemos comentado, los modelos 1 y 2 no se tienen en cuenta). En los modelos del 9 al 16, se utilizó el modelo de identificación que mejor porcentaje obtenía, Modelo 7. Los modelos 17 y 18, se obtuvieron aplicando el Modelo 3 de identificación pero no se

tuvieron en cuenta los atributos correspondientes a los aciertos y fallos indicados por el modelo de identificación.

A continuación se describen los modelos obtenidos:

- **Modelo 1:** este modelo se realizó con todas las palabras de los ficheros y con los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion, FallosIdentificacion, Atributos del Análisis y Es_entidad_corpus. Para este modelo, se obtuvieron los siguientes resultados:

Correctly Classified Instances	48717	95.8411 %
Incorrectly Classified Instances	2114	4.1589 %

Class	Precision	Recall	F-Measure
PER	0,884	0,863	0,873
ORG	0,637	0,685	0,66
LOC	0,732	0,779	0,755
MISC	0,668	0,554	0,605
NA	1	1	1

Tabla 55. Resultados Modelo 1 de Clasificación

- **Modelo 2:** en este modelo se utilizaron todas las palabras de los ficheros de entrenamiento y prueba y los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion, FallosIdentificacion, Atributos del Análisis y Es_entidad_clasificacion. Los resultados aportados por Weka a la hora de obtener este modelo fueron:

Correctly Classified Instances	47468	93.384 %
Incorrectly Classified Instances	3363	6.616 %

Class	Precision	Recall	F-Measure
PER	0,848	0,848	0,848
ORG	0,591	0,538	0,563
LOC	0,732	0,732	0,732
MISC	0,557	0,306	0,395
NA	0,971	0,989	0,98

Tabla 56. Resultados Modelo 2 de Clasificación

- **Modelo 3:** este modelo se realizó con todas las palabras de los ficheros y con los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion, FallosIdentificacion y Es_entidad_corpus. Se obtuvieron los siguientes datos:

Correctly Classified Instances	48694	95.7959 %
Incorrectly Classified Instances	2137	4.2041 %

Class	Precision	Recall	F-Measure
PER	0,88	0,865	0,872
ORG	0,626	0,69	0,657
LOC	0,742	0,787	0,764
MISC	0,657	0,509	0,574
NA	1	1	1

Tabla 57. Resultados Modelo 3 de Clasificación

- **Modelo 4:** este modelo también se obtuvo utilizando todas las palabras de los ficheros y los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig,

CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion, FallosIdentificacion y Es_entidad_clasificacion. En este modelo, se obtuvieron los siguientes datos:

Correctly Classified Instances	47395	93.2403 %
Incorrectly Classified Instances	3436	6.7597 %

Class	Precision	Recall	F-Measure
PER	0,848	0,84	0,844
ORG	0,574	0,527	0,55
LOC	0,726	0,724	0,725
MISC	0,57	0,278	0,373
NA	0,97	0,989	0,979

Tabla 58. Resultados Modelo 4 de Clasificación

- **Modelo 5:** para realizar este modelo, se utilizaron los atributos que según el corpus de entrenamiento eran entidades. Los atributos utilizados fueron: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion, FallosIdentificacion y Atributos del Análisis. Los resultados para este modelo fueron:

Correctly Classified Instances	6474	75.3843 %
Incorrectly Classified Instances	2114	24.6157 %

Class	Precision	Recall	F-Measure
PER	0,884	0,863	0,873
ORG	0,637	0,685	0,66
LOC	0,732	0,779	0,755
MISC	0,668	0,554	0,605
NA	0	0	0

Tabla 59. Resultados Modelo 5 de Clasificación

- **Modelo 6:** al contrario que en el modelo anterior, este se realizó únicamente con las palabras que según el modelo de identificación, previamente aplicado, eran entidad. Los atributos utilizados fueron: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion, FallosIdentificacion y Atributos del Análisis. Y los resultados obtenidos:

Correctly Classified Instances	6338	85.4984 %
Incorrectly Classified Instances	1075	14.5016 %

Class	Precision	Recall	F-Measure
PER	0,942	0,913	0,927
ORG	0,629	0,627	0,628
LOC	0,776	0,833	0,804
MISC	0,677	0,27	0,386
NA	0,904	0,96	0,932

Tabla 60. Resultados Modelo 6 de Clasificación

- **Modelo 7:** modelo realizado con las palabras que son entidad según el corpus de entrenamiento y con los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion y FallosIdentificacion. Se obtuvieron los siguientes datos:

Correctly Classified Instances	6451	75.1164 %
Incorrectly Classified Instances	2137	24.8836 %

Class	Precision	Recall	F-Measure
PER	0,88	0,865	0,872

ORG	0,626	0,69	0,657
LOC	0,742	0,787	0,764
MISC	0,657	0,509	0,574
NA	0	0	0

Tabla 61. Resultados Modelo 7 de Clasificación

- **Modelo 8:** modelo obtenido utilizando sólo las palabras que el modelo de identificación indicaba eran entidades y con los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion y FallosIdentificacion. Los resultados fueron:

Correctly Classified Instances	6306	85.0668 %
Incorrectly Classified Instances	1107	14.9332 %

Class	Precision	Recall	F-Measure
PER	0,94	0,909	0,924
ORG	0,633	0,601	0,617
LOC	0,774	0,821	0,797
MISC	0,701	0,229	0,345
NA	0,889	0,969	0,928

Tabla 62. Resultados Modelo 8 de Clasificación

- **Modelo 9:** este modelo se realizó con todas las palabras de los ficheros y con los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion, FallosIdentificacion, Atributos del Análisis y Es_entidad_corpus. Para este modelo, se obtuvieron los siguientes resultados:

Correctly Classified Instances	48758	95.9218 %
Incorrectly Classified Instances	2073	4.0782 %

Class	Precision	Recall	F-Measure
PER	0,875	0,87	0,872
ORG	0,642	0,696	0,668
LOC	0,75	0,781	0,766
MISC	0,677	0,547	0,605
NA	1	1	1

Tabla 63. Resultados Modelo 9 de Clasificación

- **Modelo 10:** en este modelo se utilizaron todas las palabras de los ficheros de entrenamiento y prueba y los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion, FallosIdentificacion, Atributos del Análisis y Es_entidad_clasificacion. Los resultados aportados por Weka a la hora de obtener este modelo fueron:

Correctly Classified Instances	47548	93.5413 %
Incorrectly Classified Instances	3283	6.4587 %

Class	Precision	Recall	F-Measure
PER	0,861	0,848	0,854
ORG	0,59	0,563	0,576
LOC	0,733	0,715	0,724
MISC	0,588	0,357	0,445
NA	0,973	0,989	0,981

Tabla 64. Resultados Modelo 10 de Clasificación

- **Modelo 11:** este modelo se realizó con todas las palabras de los ficheros y con los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc,

En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion, FallosIdentificacion y Es_entidad_corpus. Se obtuvieron los siguientes datos:

Correctly Classified Instances	48759	95.9237 %
Incorrectly Classified Instances	2072	4.0763 %

Class	Precision	Recall	F-Measure
PER	0,876	0,872	0,874
ORG	0,641	0,695	0,667
LOC	0,747	0,783	0,764
MISC	0,684	0,543	0,605
NA	1	1	1

Tabla 65. Resultados Modelo 11 de Clasificación

- **Modelo 12:** este modelo también se obtuvo utilizando todas las palabras de los ficheros y los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion, FallosIdentificacion y Es_entidad_clasificacion. En este modelo, se obtuvieron los siguientes datos:

Correctly Classified Instances	47521	93.4882 %
Incorrectly Classified Instances	3310	6.5118 %

Class	Precision	Recall	F-Measure
PER	0,852	0,848	0,85
ORG	0,583	0,558	0,57
LOC	0,739	0,714	0,726
MISC	0,579	0,354	0,44

NA	0,973	0,988	0,981

Tabla 66. Resultados Modelo 12 de Clasificación

- **Modelo 13:** para realizar este modelo, se utilizaron los atributos que según el corpus de entrenamiento eran entidades. Los atributos utilizados fueron: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion, FallosIdentificacion y Atributos del Análisis. Los resultados para este modelo fueron:

Correctly Classified Instances	6515	75.8617 %
Incorrectly Classified Instances	2073	24.1383 %

Class	Precision	Recall	F-Measure
PER	0,875	0,87	0,872
ORG	0,642	0,696	0,668
LOC	0,75	0,781	0,766
MISC	0,677	0,547	0,605
NA	0	0	0

Tabla 67. Resultados Modelo 13 de Clasificación

- **Modelo 14:** al contrario que en el modelo anterior, este se realizó únicamente con las palabras que según el modelo de identificación, previamente aplicado, eran entidad. Los atributos utilizados fueron: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion, FallosIdentificacion y Atributos del Análisis. Y los resultados obtenidos:

Correctly Classified Instances	7499	75.9546 %

Incorrectly Classified Instances 2374 24.0454 %

Class	Precision	Recall	F-Measure
PER	0,861	0,872	0,867
ORG	0,591	0,663	0,625
LOC	0,733	0,754	0,743
MISC	0,587	0,424	0,493
NA	0,838	0,815	0,827

Tabla 68. Resultados Modelo 14 de Clasificación

- **Modelo 15:** modelo realizado con las palabras que son entidad según el corpus de entrenamiento y con los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion y FallosIdentificacion. Se obtuvieron los siguientes datos:

Correctly Classified Instances 6516 75.8733 %
Incorrectly Classified Instances 2072 24.1267 %

Class	Precision	Recall	F-Measure
PER	0,876	0,872	0,874
ORG	0,641	0,695	0,667
LOC	0,747	0,783	0,764
MISC	0,684	0,543	0,605
NA	0	0	0

Tabla 69. Resultados Modelo 15 de Clasificación

- **Modelo 16:** modelo obtenido utilizando sólo las palabras que el modelo de identificación indicaba eran entidades y con los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig,

En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion y FallosIdentificacion. Los resultados fueron:

Correctly Classified Instances	7467	75.6305 %
Incorrectly Classified Instances	2406	24.3695 %

Class	Precision	Recall	F-Measure
PER	0,853	0,87	0,861
ORG	0,583	0,658	0,618
LOC	0,738	0,75	0,744
MISC	0,579	0,426	0,491
NA	0,84	0,811	0,826

Tabla 70. Resultados Modelo 16 de Clasificación

- **Modelo 17:** este modelo se realizó con todas las palabras de los ficheros y con los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion, FallosIdentificacion, Atributos del Análisis y Es_entidad_corpus. Para este modelo, se obtuvieron los siguientes resultados:

Correctly Classified Instances	48708	95.8234 %
Incorrectly Classified Instances	2123	4.1766 %

Class	Precision	Recall	F-Measure
PER	0,879	0,862	0,871
ORG	0,637	0,686	0,66
LOC	0,732	0,774	0,752
MISC	0,672	0,556	0,609
NA	1	1	1

Tabla 71. Resultados Modelo 17 de Clasificación

- **Modelo 18:** en este modelo se utilizaron todas las palabras de los ficheros de entrenamiento y prueba y los siguientes atributos: Case, ContieneDigitos, ContieneSimbolos, Categoria, Categoria_sig, Categoria_ant, En_listado_per, En_listado_org, En_listado_loc, En_listado_misc, En_listado_per_ant, En_listado_org_ant, En_listado_loc_ant, En_listado_misc_ant, En_listado_per_sig, En_listado_org_sig, En_listado_loc_sig, En_listado_misc_sig, CompletoEnListado, NumeroPalabrasListado, AciertosIdentificacion, FallosIdentificacion, Atributos del Análisis y Es_entidad_clasificacion. Los resultados aportados por Weka a la hora de obtener este modelo fueron:

Correctly Classified Instances	47435	93.319 %
Incorrectly Classified Instances	3396	6.681 %

Class	Precision	Recall	F-Measure
PER	0,849	0,85	0,849
ORG	0,595	0,533	0,562
LOC	0,726	0,723	0,724
MISC	0,555	0,289	0,381
NA	0,97	0,989	0,979

Tabla 72. Resultados Modelo 18 de Clasificación

6. Resultados

En este capítulo, se procederá a comentar los resultados obtenidos en los diferentes modelos que se han creado y que fueron comentados en el apartado anterior. Obviamente, se comentarán por separado los resultados correspondientes al Modelo de Identificación y los correspondientes al Modelo de Clasificación debido a que obtienen diferente información.

6.1. Modelo de Identificación

Como posibles modelos a implementar del Modelo de Identificación, se realizaron 9 modelos. La forma de evaluar dichos modelos, aparte del porcentaje de entidades correctamente clasificadas, es mediante el precision y el recall (datos especificados en el apartado anterior). La función de rendimiento con la que se trabajará fue comentada en el apartado 2.1.

En la figura 17, se puede ver un resumen de los valores obtenidos de cada uno de los modelos de identificación.

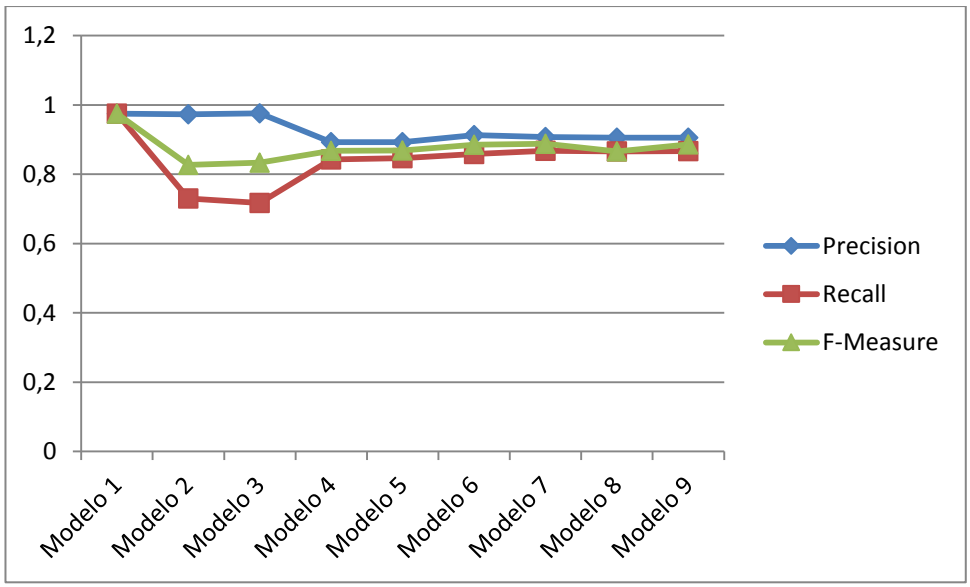

Ilustración 18. Resumen de los resultados de los Modelos de Identificación

De los 9 modelos hay que comentar lo siguiente:

- Como ya se dijo en un apartado anterior, el Modelo 1 que obtuvo un porcentaje de entidades clasificadas correctamente de 99.1403 % hay que volver a comentar que se realizó con unos listados obtenidos de los propios ficheros de entrenamiento y prueba (cada uno con sus respectivos listados), por lo que no

podemos implementar este modelo debido a que uno de los requisitos de la aplicación era que los ficheros debían ser los provistos por Yago (comentado en el apartado 3.1). Esto no quiere decir que no nos aporte información, ya que podemos decir que los listados son una herramienta muy importante para esta aplicación y que los modelos obtenidos de esta manera son correctos.

- El Modelo 2 es que modelo que obtiene el peor resultado de entre los 9. Esto es debido a que al utilizar unos listados obtenidos del fichero de entrenamiento para obtener el fichero de prueba, es bastante probable que muchas entidades de dicho fichero no se encontraran en los listados. Aun así, se obtiene un resultado bastante alto. En este modelo, si nos fijamos en el recall obtenido para entity, se puede ver que se consideran entidades a tokens que en verdad no lo son (falsos positivos).
- A parte de los dos modelos anteriores que como se comentó no se tendrían en cuenta, el Modelo 3 es el que obtiene peor resultado en cuanto a instancias clasificadas correctamente. Si nos fijamos en la función de rendimiento, quizás no se note tanta diferencia con los demás modelos, pero si nos fijamos en el recall obtenido para entity, nos damos cuenta de el gran número de falsos positivos que hubo.
- En el Modelo 4, realizado sin listados, se obtuvieron mejores resultados que con el modelo anterior. Aunque se identificaran menos tokens como *entity*, hubo más cantidad de aciertos dentro de las identificaciones (menos falsos positivos), lo que se considera mejor a la hora de evaluar un modelo.
- El Modelo 5 muestra una muy leve mejoría con respecto al Modelo 4. Esta mejoría se podría atribuir al uso de los Atributos del Análisis.
- En el Modelo 6, podemos observar una notable mejoría (teniendo en cuenta que los resultados que se obtienen son cercanos al 100%) en cuanto al número de instancias correctamente clasificadas. Esta mejoría también se hace visible debido a que se recupera un número mayor de entidades y hay menos falsos positivos. Esta diferencia con los modelos anteriores, las podemos atribuir al uso de los listados de Yago, en vez de otros listados o no usar ningún listado.
- El modelo que mejor resultados obtiene es el Modelo 7, aunque no existe una diferencia demasiado marcada con el modelo anterior. De nuevo, atribuída a la utilización de los Atributos del Análisis.

- Los modelos 8 y 9, obtienen unos resultados similares aunque ligeramente peores que los obtenidos por los modelos 6 y 7. Para obtener aquéllos modelos, se realizaron algunos cambios en los listados de Yago, ya que en los ficheros obtenidos inicialmente de la ontología Yago, había algunos errores o lo que se entendieron como errores. Cosa que da que pensar debido a que con los ficheros modificados se obtienen resultados peores que con los iniciales.

6.2. Modelo de Clasificación

Para la evaluación de los Modelos de Clasificación obtenidos, se seguirá el mismo procedimiento que para los Modelos de Identificación. Es posible que en este apartado se realicen alusiones a otros datos en cuyo caso se debería ver el Apéndice 1 que es donde se encuentran los resúmenes completos de todos los modelos realizados.

Hay que tener en cuenta que los modelos que utilizan información importante dada por el corpus utilizado, ya sea las palabras que realmente son entidad o el tipo de entidad de un token, no se podrán implementar en la aplicación. Esto es debido a que parten de información que cuando se ejecute la aplicación con un fichero de entrada, obtenido de internet o de cualquier otro sitio, esa información no se va a conocer. Se conocerá información relativa a la salida del Modelo de Identificación que se haya ejecutado previamente, pero nunca tendrá el grado de veracidad que la información de los ficheros del corpus. Los modelos a los que se hace referencia son los modelos numerados con números impares, es decir: 1, 3, 5, 7, 9, 11, 13, 15 y 17.

La forma de evaluar estos modelos no será simple debido a que al tener varias posibilidades para clasificar una entidad, hay que tener en cuenta la función de medida de rendimiento (F) de cada posible clase (PER, ORG, LOC, MISC, NA), así como el precision y el recall de cada uno de ellos. También se tendrán en cuenta otros valores que como se dijo antes se pueden ver en el Apéndice 1.

Como se comentó en el capítulo anterior, hay que tener en cuenta que los Modelos de Clasificación se obtuvieron después de aplicar el Modelo de Identificación por lo que pueden existir errores previos.

En la figura 18, se puede ver un resumen de la media aritmética para todas las clases de cada uno de los valores obtenidos de cada uno de los modelos.

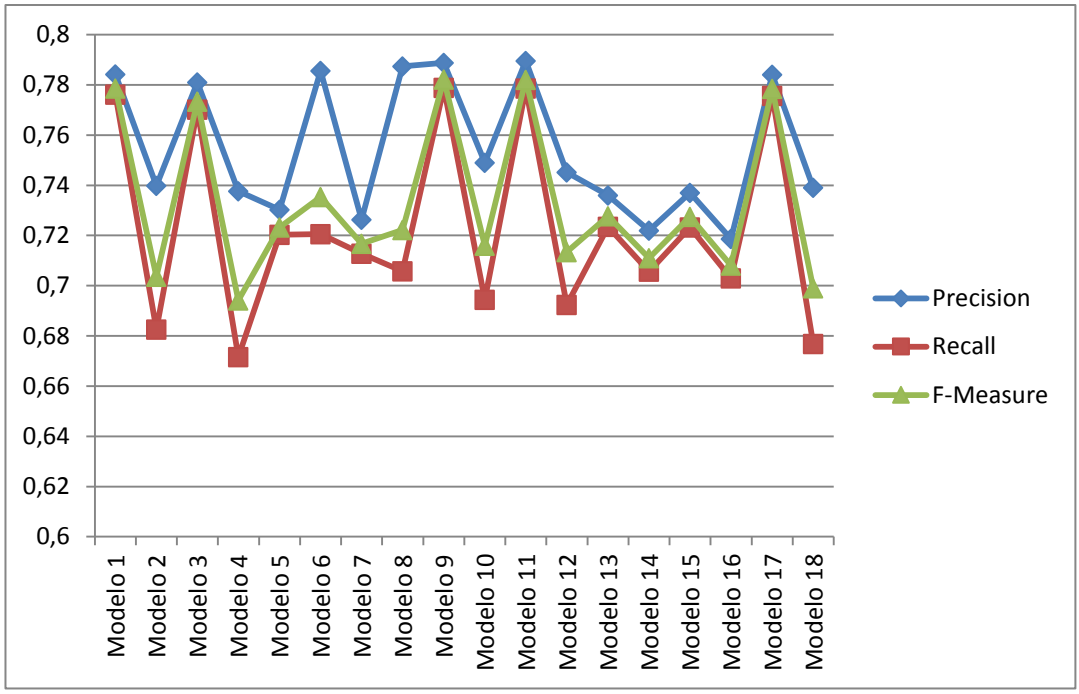

Ilustración 19. Resumen de los resultados de los Modelos de Clasificación

A continuación, se procede a comentar los resultados de cada modelo de clasificación:

- El Modelo 1 es el que mejor resultado obtiene con respecto a que etiqueta un número mayor de entidades que los otros modelos. Si nos fijamos en los valores de F-Measure, podemos ver que no es el que tiene los valores más altos para cada clase ya que, por ejemplo, el Modelo 6 tiene valores mayores para las clases PER y LOC. Aun así hay que tener en cuenta que el Modelo 6 únicamente contiene las palabras que según el Modelo de Identificación previamente utilizado (en este caso el Modelo 3) son entidad, por lo que posee un número mucho menor de instancias que el Modelo 1. Con esto, se quiere decir que no se pueden comparar modelos tan diferentes porque aunque la precision o el recall sean mayores en un modelo, el otro puede etiquetar correctamente un número mayor de entidades al tener más entidades para etiquetar.
- En el Modelo 2 se puede ver cómo influye el Modelo de Identificación aplicado, debido a que se reconocen menos entidades. Aun así, el porcentaje de instancias correctamente clasificadas sigue siendo bastante algo. Únicamente un 2.5% menos que en el modelo anterior.
- El Modelo 3 reconoce casi el mismo porcentaje de entidades correctamente, aunque si nos fijamos en la matriz de confusión (Apéndice 1), podemos observar

que detecta más entidades de tipo PER (4), LOC (9) y ORG (17) que el Modelo 1. Sin embargo, detecta menos entidades de tipo MISC (56) lo que hace que el número de entidades totales sea mayor.

- En el Modelo 4 se puede observar una leve bajada del número total de aciertos así como de los aciertos de cada clase de manera individual, con respecto al Modelo 2 (ya que el Modelo 1 y el Modelo 3 se realizaron con valores diferentes). Esta bajada podría asociarse a que en este modelo no se utilizan los Atributos del Análisis.
- Si nos fijamos en los resultados del Modelo 5 podemos ver que son iguales a los resultados obtenidos en el Modelo 1, a excepción del correspondiente a NA, debido a que en el Modelo 5 únicamente hemos utilizado las entidades etiquetadas en el corpus por lo que todas los tokens tratados eran entidades. En este modelo, se nota una bajada importante en el porcentaje de instancias clasificadas correctamente aun tratando con menor número de entidades.
- El Modelo 6, como se indicó al hablar del Modelo 1, obtenía valores de F-Measure mejores que otros modelos pero tenía un porcentaje más bajo de instancias clasificadas correctamente. Ya que únicamente posee los tokens que según el Modelo de Identificación aplicado son entidades, habrá entidades que se hayan perdido por el camino y tokens clasificados como entidades que en realidad no lo son.
- El Modelo 7 tiene los mismos valores que el Modelo 3 aunque tenga un porcentaje de acierto (instancias correctamente clasificadas) mucho más bajo. Esto es porque al tener menos ocurrencias, al no tenerse en cuenta los tokens que no son entidad, y al tener el mismo número de aciertos y fallos, los porcentajes globales son diferentes aunque se hayan detectado las mismas entidades y se haya fallado en las mismas entidades que el otro modelo.
- El Modelo 8 tienes unos valores muy parecidos a los del Modelo 6. La diferencia entre ambos modelos es que el Modelo 8 no utiliza los Atributos del Análisis y por ello obtiene valores menores a la hora de reconocer cada tipo de entidad, siendo más remarcado a la hora de reconocer entidades de tipo MISC.
- En el Modelo 9, que se realizó con las mismas características que el Modelo 1, se puede observar una levísima mejoría reconociendo mayor número de entidades de tipo PER, LOC y ORG, pero se puede ver que reconoce menor

número de entidades MISC aunque el número de instancias correctamente clasificadas aumenta de manera simbólica.

- El Modelo 10 mejora los valores obtenidos por su gemelo el Modelo 2 para las entidades de ORG y MISC pero empeora el resultado para entidades de tipo LOC. Aun empeorando los resultados para entidades de tipo LOC, etiqueta más entidades correctamente que el modelo con el que se le compara.
- En el Modelo 11, se puede ver que obtiene los mismos valores o mejores que el Modelo 3, dato que se puede asociar al Modelo de Identificación que se utiliza. El problema es que no siempre que se ha utilizado un Modelo de Identificación mejor, se mejoran los valores individuales para cada clase y, por extensión, los valores totales.
- El Modelo 12 mejora los datos del Modelo 4 para todas las clases, salvo para NA. En este caso, se tendría que evaluar lo que tiene más importancia: etiquetar mayor número de instancias o disminuir el número de instancias etiquetadas erróneamente.
- De nuevo, en el Modelo 13 podemos ver una notable mejoría con respecto al Modelo 5 que se atribuye a la utilización de un Modelo de Identificación diferente. Lo que también hay que tener en cuenta es que reconoce un número menor de entidades MISC (únicamente 6) aunque aumenta el mayor número para el resto de clases.
- Aunque el Modelo 14 tiene un porcentaje de clasificación de instancias correctamente considerablemente menor que el Modelo 6, el número de entidades clasificadas correctamente es mayor debido a que al utilizar un Modelo de Identificación mejor, se ha detectado mayor número de entidades. Además de tener mayor número de tokens identificados como *entity,* existe un menor número de *'noentity'* (identificadas por el algoritmo como *'entity'*) que en el Modelo 6. Lo único malo es que se etiqueta como NA un número mayor de entidades de otras clases que en el Modelo 6, aunque por probabilidad puede pasar esto ya que, se trabaja con un número mayor de instancias.
- En el Modelo 15, se mejoran los datos obtenidos en el Modelo 7 para todas las clases, sin tener en cuenta la clase NA, ya que en el Modelo 15 únicamente se encuentran las entidades según el corpus.
- Con el Modelo 16 y 8 pasa lo mismo que con el Modelo 14 y 6. Se detectan más entidades con un número mayor de instancias.

- El Modelo 17 mejora los resultados del Modelo 1 para MISC y ORG empeorando para LOC y PER. Al compararlo con el Modelo 9, vemos que únicamente mejora los resultados para MISC.
- El Modelo 18, mejora los resultados del Modelo 10 para LOC y PER y los del Modelo 2 para PER y NA, aunque obtiene un porcentaje de acierto menor a ambos.

7. Resumen del proyecto

Como se ha podido ver a lo largo de este proyecto, se han ido resolviendo todos los objetivos que se pusieron en su inicio. También se han comentado los pasos que se han seguido hasta obtener la herramienta actual, así como los problemas que se encontraron durante su desarrollo.

Una cosa que no se ha comentado y que es necesario comentar en este punto, es lo aprendido durante la realización del proyecto. Por un lado y creo que lo más importante, son los errores que se han cometido y que han ayudado a cambiar de puntos de vista y a aprender más, en mi opinión, que cuando se han hecho las cosas bien. Por otro lado, se ha aprendido de un dominio del que no se conocía prácticamente nada, salvo alguna asignatura cursada durante la carrera. Además del dominio, también se han utilizado aplicaciones y recursos que han resultado interesantes y, tanto el dominio como las aplicaciones, han hecho surgir mucho interés en otros posibles usos de las herramientas y en querer conocer más acerca de todo lo que conlleva la rama de la Extracción de Información y de la Recuperación de la Información, aunque en concreto del tema relacionado con las herramientas NER.

Algo que se puede echar en falta en el desarrollo de un proyecto es la planificación seguida para su desarrollo. Hay que decir que no se ha incluído debido a que según la forma en la que se ha desarrollado y todo el tiempo que me ha costado llegar hasta este punto, no hay una planificación clara. Obviamente, éste ha sido uno de los grandes errores cometidos en este desarrollo y que confío en que no volverá a suceder, al menos de una forma tan clara. Creo totalmente necesario fijarse unos límites (hitos) en el desarrollo de cualquier proyecto, ya sea un proyecto software o un proyecto de la vida personal.

8. Conclusiones

Según lo que se ha visto, puede que un modelo concreto etiquete mayor número de entidades que otro pero puede ser que si nos fijamos en el número total de entidades de cada tipo de entidad etiquetadas correctamente, un modelo etiquete mejor una clase de entidad concreta aunque etiquete menos entidades en general.

Con los modelos realizados, se ha podido comprobar que la utilización de los Atributos del Análisis mejora el número total de clasificaciones correctas aunque se tendrá que tener en cuenta que el uso de un mayor número de atributos a la hora de utilizar un modelo, influirá en el tiempo de ejecución y más cuando se trata de los Atributos del Análisis que se calculan según se ejecuta el modelo.

Se ha podido comprobar que se obtienen mejores resultados utilizando atributos relacionados con el corpus, pero no han resultado ser tan claves como se pensaba en un principio. En cualquier caso, han servido para establecer la importancia de unos u otros atributos.

La aplicación previa de un Modelo de Identificación u otro influye en el número de instancias clasificadas correctamente, aunque no tiene porqué influir a la hora de comprobar los datos de cada clase.

La utilización de unos listados u otros a la hora de obtener el Modelo de Identificación también se ha visto que influye aunque obviamente depende de qué tipo de listados se vayan a comparar. Se dice esto, ya que se obtuvieron mejores resultados sin utilizar listados que utilizando los listados obtenidos del fichero de entrenamiento. Esto quiere decir que los listados no influyen únicamente para bien, sino que también pueden influir para mal, es decir, pueden aportar información errónea.

Según se ha visto con la utilización de dos Modelos de Identificación diferentes, se puede ver que aunque un modelo obtenga mejores resultados, se pueden clasificar entidades de una clase o varias clases habiendo utilizado un modelo con peores resultados.

9. Líneas futuras

La primera posibilidad para desarrollar una nueva versión de la aplicación, sería realizar un nuevo estudio con diferentes atributos y comprobar cuáles de ellos inciden más en los resultados finales de identificación y clasificación.

Como se ha comprobado el uso de los corpus influye en el modelo que se obtiene y en los resultados, por lo que se podrían realizar pruebas con corpus diferentes a los utilizados en este caso, para mejorar los resultados y ser capaces de detectar un número mayor de entidades.

Otro punto que resultaría interesante, sería utilizar Internet en lugar de listados, de modo que se consultara la información en el momento en todo Internet o en diferentes bases de datos como puede ser Dbpedia a través del lenguaje de consultar sparql.

En el caso de que se quisiera hacer un cambio de mayor envergadura, se podrían utilizar otros métodos de aprendizaje e intentar conseguir una aplicación independiente del lenguaje, ya que como se ha dicho el corpus utilizado en este proyecto, es para el idioma inglés.

10. Bibliografía

[Odon de Alencar, Rafael, 2009]. Seminário de Gerência de Dados da Web.

[Nadeau, David, 2007]. Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision.

[Payam Refaeilzadeh et al., 2008]. Cross-validation.

[T. Mitchell, 1997]. Decision Tree Learning - Based on “Machine Learning".

[Erik F. Tjong Kim Sang and Fien De Meulder]. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.

[Fernández, Óscar et al., 2005] Procesamiento del Lenguaje Natural, núm. 35, pp. 37-44.

.

11. Anexo 1. Resúmenes de Modelos de Identificación

En ese apartado, se mostrarán los resúmenes de todos los modelos de identificación creados. Estos resúmenes, los proporciona Weka como salida a la vez que el modelo concreto.

Modelo 1

```
Correctly Classified Instances       50394               99.1403 %
Incorrectly Classified Instances       437                0.8597 %
Kappa statistic                          0.9694
Mean absolute error                      0.0177
Root mean squared error                  0.0941
Relative absolute error                  6.2969 %
Root relative squared error             25.101  %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.975     0.005      0.975     0.975     0.975      0.99     entity
                 0.995     0.025      0.995     0.995     0.995      0.99     noentity
Weighted Avg.    0.991     0.022      0.991     0.991     0.991      0.99

=== Confusion Matrix ===

     a     b   <-- classified as
  8370   218 |     a = entity
   219 42024 |     b = noentity
```

Ilustración 20. Resultados Modelo 1 de Identificación

Modelo 2

```
Correctly Classified Instances       48249               94.9204 %
Incorrectly Classified Instances      2582                5.0796 %
Kappa statistic                          0.7978
Mean absolute error                      0.0556
Root mean squared error                  0.2173
Relative absolute error                 19.8168 %
Root relative squared error             57.9965 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.717     0.004      0.976     0.717     0.827      0.953    entity
                 0.996     0.283      0.945     0.996     0.97       0.953    noentity
Weighted Avg.    0.949     0.236      0.951     0.949     0.946      0.953

=== Confusion Matrix ===

     a     b   <-- classified as
  6160  2428 |     a = entity
   154 42089 |     b = noentity
```

Ilustración 21. Resultados Modelo 2 de Identificación

Modelo 3

```
Correctly Classified Instances       48341               95.1014 %
Incorrectly Classified Instances      2490                4.8986 %
Kappa statistic                          0.8063
Mean absolute error                      0.0557
Root mean squared error                  0.2173
Relative absolute error                 19.8663 %
Root relative squared error             58.0048 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.73      0.004      0.973      0.73      0.834      0.953    entity
                 0.996     0.27       0.948      0.996     0.971      0.953    noentity
Weighted Avg.    0.951     0.225      0.952      0.951     0.948      0.953

=== Confusion Matrix ===

     a     b   <-- classified as
  6269  2319 |     a = entity
   171 42072 |     b = noentity
```

Ilustración 22. Resultados Modelo 3 de Identificación

Modelo 4

```
Correctly Classified Instances       48619               95.6483 %
Incorrectly Classified Instances      2212                4.3517 %
Kappa statistic                          0.8415
Mean absolute error                      0.0735
Root mean squared error                  0.1929
Relative absolute error                 26.2106 %
Root relative squared error             51.493  %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.843     0.021      0.893      0.843     0.868      0.96     entity
                 0.979     0.157      0.969      0.979     0.974      0.96     noentity
Weighted Avg.    0.956     0.134      0.956      0.956     0.956      0.96

=== Confusion Matrix ===

     a     b   <-- classified as
  7243  1345 |     a = entity
   867 41376 |     b = noentity
```

Ilustración 23. Resultados Modelo 4 de Identificación

Modelo 5

```
Correctly Classified Instances       48648               95.7054 %
Incorrectly Classified Instances      2183                4.2946 %
Kappa statistic                          0.8438
Mean absolute error                      0.0724
Root mean squared error                  0.1925
Relative absolute error                 25.7972 %
Root relative squared error             51.3733 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.847     0.021      0.893     0.847     0.869      0.959    entity
                 0.979     0.153      0.969     0.979     0.974      0.959    noentity
Weighted Avg.    0.957     0.131      0.956     0.957     0.957      0.959

=== Confusion Matrix ===

     a     b   <-- classified as
  7272  1316 |     a = entity
   867 41376 |     b = noentity
```

Ilustración 24. Resultados Modelo 5 de Identificación

Modelo 6

```
Correctly Classified Instances       48914               96.2287 %
Incorrectly Classified Instances      1917                3.7713 %
Kappa statistic                          0.8625
Mean absolute error                      0.0605
Root mean squared error                  0.1772
Relative absolute error                 21.5808 %
Root relative squared error             47.2832 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.859     0.017      0.913     0.859     0.885      0.97     entity
                 0.983     0.141      0.972     0.983     0.977      0.97     noentity
Weighted Avg.    0.962     0.12       0.962     0.962     0.962      0.97

=== Confusion Matrix ===

     a     b   <-- classified as
  7376  1212 |     a = entity
   705 41538 |     b = noentity
```

Ilustración 25. Resultados Modelo 6 de Identificación

Modelo 7

```
Correctly Classified Instances       48942               96.2838 %
Incorrectly Classified Instances      1889                3.7162 %
Kappa statistic                          0.8653
Mean absolute error                      0.0603
Root mean squared error                  0.1776
Relative absolute error                 21.5034 %
Root relative squared error             47.3999 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.868     0.018      0.908     0.868     0.888      0.97     entity
                 0.982     0.132      0.973     0.982     0.978      0.97     noentity
Weighted Avg.    0.963     0.113      0.962     0.963     0.963      0.97

=== Confusion Matrix ===

     a     b   <-- classified as
  7456  1132 |     a = entity
   757 41486 |     b = noentity
```

Ilustración 26. Resultados Modelo 7 de Identificación

Modelo 8

```
Correctly Classified Instances       48908               96.2169 %
Incorrectly Classified Instances      1923                3.7831 %
Kappa statistic                          0.8629
Mean absolute error                      0.062
Root mean squared error                  0.1773
Relative absolute error                 22.1029 %
Root relative squared error             47.3199 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.866     0.018      0.906     0.866     0.886      0.972    entity
                 0.982     0.134      0.973     0.982     0.977      0.972    noentity
Weighted Avg.    0.962     0.114      0.962     0.962     0.962      0.972

=== Confusion Matrix ===

     a     b   <-- classified as
  7441  1147 |     a = entity
   776 41467 |     b = noentity
```

Ilustración 27. Resultados Modelo 8 de Identificación

Modelo 9

```
Correctly Classified Instances       48919               96.2385 %
Incorrectly Classified Instances      1912                3.7615 %
Kappa statistic                          0.8637
Mean absolute error                      0.0609
Root mean squared error                  0.1771
Relative absolute error                 21.7028 %
Root relative squared error             47.2626 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.867     0.018      0.906     0.867     0.886      0.97     entity
                 0.982     0.133      0.973     0.982     0.977      0.97     noentity
Weighted Avg.    0.962     0.113      0.962     0.962     0.962      0.97

=== Confusion Matrix ===

     a     b   <-- classified as
  7449  1139 |     a = entity
   773 41470 |     b = noentity
```

Ilustración 28. Resultados Modelo 9 de Identificación

12. Anexo 2. Resúmenes de Modelos de Clasificación

En ese apartado, se mostrarán los resúmenes de todos los modelos de clasificación creados. Estos resúmenes, los proporciona Weka como salida a la vez que el modelo concreto.

Modelo 1

```
Correctly Classified Instances       48717               95.8411 %
Incorrectly Classified Instances      2114                4.1589 %
Kappa statistic                          0.8621
Mean absolute error                      0.0219
Root mean squared error                  0.111
Relative absolute error                 18.2087 %
Root relative squared error             45.1891 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.863     0.007       0.884     0.863     0.873      0.993    PER
  0.685     0.017       0.637     0.685     0.66       0.976    ORG
  0.779     0.012       0.732     0.779     0.755      0.979    LOC
  0.554     0.007       0.668     0.554     0.605      0.956    MISC
  1         0           1         1         1          1        NA

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
  2718   199   182    50     0 |     a = PER
   157  1425   293   205     0 |     b = ORG
   107   261  1629    94     0 |     c = LOC
    93   352   121   702     0 |     d = MISC
     0     0     0     0 42243 |     e = NA
```

Ilustración 29. Resultados Modelo 1 de Clasificación

Modelo 2

```
Correctly Classified Instances       47468               93.384  %
Incorrectly Classified Instances      3363                6.616  %
Kappa statistic                          0.7715
Mean absolute error                      0.0384
Root mean squared error                  0.1439
Relative absolute error                 31.824  %
Root relative squared error             58.5911 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.848     0.01       0.848     0.848     0.848      0.982    PER
  0.538     0.016      0.591     0.538     0.563      0.917    ORG
  0.732     0.011      0.732     0.732     0.732      0.955    LOC
  0.306     0.006      0.557     0.306     0.395      0.869    MISC
  0.989     0.145      0.971     0.989     0.98       0.972    NA

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
  2670   144   142    39   154 |     a = PER
   156  1119   239   119   447 |     b = ORG
   110   191  1530    70   190 |     c = LOC
    89   258    82   388   451 |     d = MISC
   123   183    96    80 41761 |     e = NA
```

Ilustración 30. Resultados Modelo 2 de Clasificación

Modelo 3

```
Correctly Classified Instances       48694               95.7959 %
Incorrectly Classified Instances      2137                4.2041 %
Kappa statistic                          0.8605
Mean absolute error                      0.0226
Root mean squared error                  0.1117
Relative absolute error                 18.719  %
Root relative squared error             45.4642 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.865     0.008      0.88      0.865     0.872      0.993    PER
  0.69      0.018      0.626     0.69      0.657      0.977    ORG
  0.787     0.012      0.742     0.787     0.764      0.979    LOC
  0.509     0.007      0.657     0.509     0.574      0.957    MISC
  1         0          1         1         1          1        NA

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
  2723   198   170    58     0 |     a = PER
   159  1436   288   197     0 |     b = ORG
   114   249  1646    82     0 |     c = LOC
    98   410   114   646     0 |     d = MISC
     0     0     0     0 42243 |     e = NA
```

Ilustración 31. Resultados Modelo 3 de Clasificación

Modelo 4

```
Correctly Classified Instances       47395               93.2403 %
Incorrectly Classified Instances      3436                6.7597 %
Kappa statistic                          0.7654
Mean absolute error                      0.0395
Root mean squared error                  0.1457
Relative absolute error                 32.8069 %
Root relative squared error             59.298  %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.84      0.01       0.848     0.84      0.844      0.981    PER
  0.527     0.017      0.574     0.527     0.55       0.919    ORG
  0.724     0.012      0.726     0.724     0.725      0.955    LOC
  0.278     0.005      0.57      0.278     0.373      0.885    MISC
  0.989     0.152      0.97      0.989     0.979      0.973    NA

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
  2646   151   148    38   166 |     a = PER
   159  1097   256   107   461 |     b = ORG
   109   216  1514    52   200 |     c = LOC
    85   282    68   352   481 |     d = MISC
   122   166   100    69 41786 |     e = NA
```

Ilustración 32. Resultados Modelo 4 de Clasificación

Modelo 5

```
Correctly Classified Instances        6474               75.3843 %
Incorrectly Classified Instances      2114               24.6157 %
Kappa statistic                          0.6602
Mean absolute error                      0.1299
Root mean squared error                  0.27
Relative absolute error                 44.5191 %
Root relative squared error             70.6638 %
Total Number of Instances             8588

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.863     0.066      0.884     0.863     0.873      0.959    PER
  0.685     0.125      0.637     0.685     0.66       0.859    ORG
  0.779     0.092      0.732     0.779     0.755      0.909    LOC
  0.554     0.048      0.668     0.554     0.605      0.861    MISC
  0         0          0         0         0          ?        NA

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 2718  199  182   50    0 |    a = PER
  157 1425  293  205    0 |    b = ORG
  107  261 1629   94    0 |    c = LOC
   93  352  121  702    0 |    d = MISC
    0    0    0    0    0 |    e = NA
```

Ilustración 33. Resultados Modelo 5 de Clasificación

Modelo 6

```
Correctly Classified Instances        6338               85.4984 %
Incorrectly Classified Instances      1075               14.5016 %
Kappa statistic                          0.7927
Mean absolute error                      0.0828
Root mean squared error                  0.214
Relative absolute error                 29.1034 %
Root relative squared error             56.7184 %
Total Number of Instances             7413

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.913     0.017      0.942     0.913     0.927      0.988    PER
  0.627     0.047      0.629     0.627     0.628      0.895    ORG
  0.833     0.046      0.776     0.833     0.804      0.952    LOC
  0.27      0.007      0.677     0.27      0.386      0.811    MISC
  0.96      0.08       0.904     0.96      0.932      0.965    NA

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1583   45   77    4   25 |    a = PER
   35  526  146   11  121 |    b = ORG
   35  105  989   12   46 |    c = LOC
   16   95   33  105  140 |    d = MISC
   12   65   29   23 3135 |    e = NA
```

Ilustración 34. Resultados Modelo 6 de Clasificación

Modelo 7

```
Correctly Classified Instances        6451               75.1164 %
Incorrectly Classified Instances      2137               24.8836 %
Kappa statistic                          0.656
Mean absolute error                      0.1336
Root mean squared error                  0.2717
Relative absolute error                 45.7668 %
Root relative squared error             71.0941 %
Total Number of Instances             8588

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.865     0.068      0.88      0.865     0.872      0.957    PER
  0.69      0.132      0.626     0.69      0.657      0.857    ORG
  0.787     0.088      0.742     0.787     0.764      0.91     LOC
  0.509     0.046      0.657     0.509     0.574      0.859    MISC
  0         0          0         0         0          ?        NA

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 2723  198  170   58    0 |    a = PER
  159 1436  288  197    0 |    b = ORG
  114  249 1646   82    0 |    c = LOC
   98  410  114  646    0 |    d = MISC
    0    0    0    0    0 |    e = NA
```

Ilustración 35. Resultados Modelo 7 de Clasificación

Modelo 8

```
Correctly Classified Instances        6306               85.0668 %
Incorrectly Classified Instances      1107               14.9332 %
Kappa statistic                          0.7853
Mean absolute error                      0.0868
Root mean squared error                  0.2181
Relative absolute error                 30.5045 %
Root relative squared error             57.807  %
Total Number of Instances             7413

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.909     0.018      0.94      0.909     0.924      0.987    PER
  0.601     0.044      0.633     0.601     0.617      0.891    ORG
  0.821     0.046      0.774     0.821     0.797      0.949    LOC
  0.229     0.005      0.701     0.229     0.345      0.834    MISC
  0.969     0.095      0.889     0.969     0.928      0.965    NA

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1576   42   83    4   29 |    a = PER
   34  504  151    7  143 |    b = ORG
   39  110  974   10   54 |    c = LOC
   16   92   25   89  167 |    d = MISC
   11   48   25   17 3163 |    e = NA
```

Ilustración 36. Resultados Modelo 8 de Clasificación

Modelo 9

```
Correctly Classified Instances       48758               95.9218 %
Incorrectly Classified Instances      2073                4.0782 %
Kappa statistic                          0.8647
Mean absolute error                      0.0215
Root mean squared error                  0.111
Relative absolute error                 17.8087 %
Root relative squared error             45.2077 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.87      0.008      0.875      0.87      0.872      0.991    PER
  0.696     0.017      0.642      0.696     0.668      0.972    ORG
  0.781     0.011      0.75       0.781     0.766      0.976    LOC
  0.547     0.007      0.677      0.547     0.605      0.949    MISC
  1         0          1          1         1          1        NA

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
  2739   190   163    57     0 |     a = PER
   166  1448   281   185     0 |     b = ORG
   124   244  1634    89     0 |     c = LOC
   101   373   100   694     0 |     d = MISC
     0     0     0     0 42243 |     e = NA
```

Ilustración 37. Resultados Modelo 9 de Clasificación

Modelo 10

```
Correctly Classified Instances       47548               93.5413 %
Incorrectly Classified Instances      3283                6.4587 %
Kappa statistic                          0.7777
Mean absolute error                      0.0367
Root mean squared error                  0.1431
Relative absolute error                 30.4803 %
Root relative squared error             58.2528 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.848     0.009      0.861     0.848     0.854      0.977    PER
  0.563     0.017      0.59      0.563     0.576      0.905    ORG
  0.715     0.011      0.733     0.715     0.724      0.946    LOC
  0.357     0.006      0.588     0.357     0.445      0.866    MISC
  0.989     0.137      0.973     0.989     0.981      0.97     NA

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
  2669   157   134    37   152 |     a = PER
   139  1170   225   115   431 |     b = ORG
    99   217  1495    63   217 |     c = LOC
    81   262    94   453   378 |     d = MISC
   113   176    91   102 41761 |     e = NA
```

Ilustración 38. Resultados Modelo 10 de Clasificación

Modelo 11

```
Correctly Classified Instances       48759               95.9237 %
Incorrectly Classified Instances      2072                4.0763 %
Kappa statistic                          0.8648
Mean absolute error                      0.0219
Root mean squared error                  0.1107
Relative absolute error                 18.2046 %
Root relative squared error             45.0536 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.872     0.008       0.876     0.872     0.874      0.991    PER
  0.695     0.017       0.641     0.695     0.667      0.972    ORG
  0.783     0.011       0.747     0.783     0.764      0.977    LOC
  0.543     0.006       0.684     0.543     0.605      0.955    MISC
  1         0           1         1         1          1        NA

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
  2745   189   159    56     0 |     a = PER
   168  1446   290   176     0 |     b = ORG
   126   242  1637    86     0 |     c = LOC
    96   378   106   688     0 |     d = MISC
     0     0     0     0 42243 |     e = NA
```

Ilustración 39. Resultados Modelo 11 de Clasificación

Modelo 12

```
Correctly Classified Instances       47521               93.4882 %
Incorrectly Classified Instances      3310                6.5118 %
Kappa statistic                          0.7762
Mean absolute error                      0.0374
Root mean squared error                  0.1438
Relative absolute error                 30.9973 %
Root relative squared error             58.5513 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.848     0.01       0.852    0.848      0.85      0.976    PER
  0.558     0.017      0.583    0.558      0.57      0.901    ORG
  0.714     0.011      0.739    0.714      0.726     0.944    LOC
  0.354     0.007      0.579    0.354      0.44      0.857    MISC
  0.988     0.136      0.973    0.988      0.981     0.971    NA

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
  2670   174   136    34   135 |     a = PER
   142  1160   227   117   434 |     b = ORG
   108   216  1493    65   209 |     c = LOC
    88   270    75   449   386 |     d = MISC
   124   171    89   110 41749 |     e = NA
```

Ilustración 40. Resultados Modelo 12 de Clasificación

Modelo 13

```
Correctly Classified Instances        6515               75.8617 %
Incorrectly Classified Instances      2073               24.1383 %
Kappa statistic                          0.6663
Mean absolute error                      0.1271
Root mean squared error                  0.2701
Relative absolute error                 43.5411 %
Root relative squared error             70.6929 %
Total Number of Instances             8588

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.87      0.072      0.875      0.87      0.872      0.955    PER
  0.696     0.124      0.642      0.696     0.668      0.86     ORG
  0.781     0.084      0.75       0.781     0.766      0.906    LOC
  0.547     0.045      0.677      0.547     0.605      0.858    MISC
  0         0          0          0         0          ?        NA

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 2739  190  163   57    0 |    a = PER
  166 1448  281  185    0 |    b = ORG
  124  244 1634   89    0 |    c = LOC
  101  373  100  694    0 |    d = MISC
    0    0    0    0    0 |    e = NA
```

Ilustración 41. Resultados Modelo 13 de Clasificación

Modelo 14

```
Correctly Classified Instances        7499               75.9546 %
Incorrectly Classified Instances      2374               24.0454 %
Kappa statistic                          0.6886
Mean absolute error                      0.1255
Root mean squared error                  0.2666
Relative absolute error                 40.2886 %
Root relative squared error             67.5939 %
Total Number of Instances             9873

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.872     0.062      0.861     0.872     0.867      0.964    PER
  0.663     0.097      0.591     0.663     0.625      0.872    ORG
  0.754     0.067      0.733     0.754     0.743      0.91     LOC
  0.424     0.029      0.587     0.424     0.493      0.843    MISC
  0.815     0.048      0.838     0.815     0.827      0.938    NA

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 2632  153  134   33   65 |    a = PER
  137 1143  218  110  115 |    b = ORG
   99  212 1465   58  109 |    c = LOC
   80  256   92  371   75 |    d = MISC
  109  170   89   60 1888 |    e = NA
```

Ilustración 42. Resultados Modelo 14 de Clasificación

Modelo 15

```
Correctly Classified Instances        6516               75.8733 %
Incorrectly Classified Instances      2072               24.1267 %
Kappa statistic                          0.6663
Mean absolute error                      0.1299
Root mean squared error                  0.2692
Relative absolute error                 44.509  %
Root relative squared error             70.452  %
Total Number of Instances             8588

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.872     0.072      0.876     0.872      0.874      0.955    PER
  0.695     0.124      0.641     0.695      0.667      0.856    ORG
  0.783     0.085      0.747     0.783      0.764      0.91     LOC
  0.543     0.043      0.684     0.543      0.605      0.865    MISC
  0         0          0         0          0          ?        NA

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 2745  189  159   56    0 |    a = PER
  168 1446  290  176    0 |    b = ORG
  126  242 1637   86    0 |    c = LOC
   96  378  106  688    0 |    d = MISC
    0    0    0    0    0 |    e = NA
```

Ilustración 43. Resultados Modelo 15 de Clasificación

Modelo 16

```
Correctly Classified Instances        7467               75.6305 %
Incorrectly Classified Instances      2406               24.3695 %
Kappa statistic                          0.6844
Mean absolute error                      0.1295
Root mean squared error                  0.2696
Relative absolute error                 41.5879 %
Root relative squared error             68.3555 %
Total Number of Instances             9873

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.87      0.066      0.853      0.87      0.861      0.961    PER
  0.658     0.1        0.583      0.658     0.618      0.862    ORG
  0.75      0.065      0.738      0.75      0.744      0.909    LOC
  0.426     0.03       0.579      0.426     0.491      0.832    MISC
  0.811     0.047      0.84       0.811     0.826      0.936    NA

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 2624  170  136   30   57 |    a = PER
  140 1134  220  112  117 |    b = ORG
  108  211 1458   59  107 |    c = LOC
   87  265   74  372   76 |    d = MISC
  116  165   87   69 1879 |    e = NA
```

Ilustración 44. Resultados Modelo 16 de Clasificación

Modelo 17

```
Correctly Classified Instances       48708               95.8234 %
Incorrectly Classified Instances      2123                4.1766 %
Kappa statistic                          0.8615
Mean absolute error                      0.022
Root mean squared error                  0.1111
Relative absolute error                 18.2436 %
Root relative squared error             45.2406 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.862     0.008      0.879     0.862     0.871      0.993    PER
  0.686     0.017      0.637     0.686     0.66       0.974    ORG
  0.774     0.012      0.732     0.774     0.752      0.98     LOC
  0.556     0.007      0.672     0.556     0.609      0.955    MISC
  1         0          1         1         1          1        NA

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
  2716   209   175    49     0 |     a = PER
   164  1426   294   196     0 |     b = ORG
   107   267  1618    99     0 |     c = LOC
   102   337   124   705     0 |     d = MISC
     0     0     0     0 42243 |     e = NA
```

Ilustración 45. Resultados Modelo 17 de Clasificación

Modelo 18

```
Correctly Classified Instances       47435               93.319  %
Incorrectly Classified Instances      3396                6.681  %
Kappa statistic                          0.7684
Mean absolute error                      0.0389
Root mean squared error                  0.1447
Relative absolute error                 32.2486 %
Root relative squared error             58.9276 %
Total Number of Instances            50831

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.85      0.01       0.849     0.85      0.849      0.982    PER
  0.533     0.015      0.595     0.533     0.562      0.918    ORG
  0.723     0.012      0.726     0.723     0.724      0.953    LOC
  0.289     0.006      0.555     0.289     0.381      0.875    MISC
  0.989     0.151      0.97      0.989     0.979      0.972    NA

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
  2676   142   145    38   148 |     a = PER
   154  1108   239   117   462 |     b = ORG
   110   187  1512    71   211 |     c = LOC
    89   254    80   367   478 |     d = MISC
   124   171   108    68 41772 |     e = NA
```

Ilustración 46. Resultados Modelo 18 de Clasificación